Bayesian modelling of flexible marked point processes with applications to event sequences from association football

SANTHOSH NARAYANAN

DOCTOR OF PHILOSOPHY IN STATISTICS

DEPARTMENT OF STATISTICS

UNIVERSITY OF WARWICK

MAY 2020

Contents

List of Tables **v**

List of Figures **ix**

Acknowledgements **xiii**

Declaration **xv**

Abstract **xvii**

Nomenclature **xix**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem statement and research goals 3

1.3 Theoretical foundation 5

1.4 Research methodology 6

1.5 Review of event sequence analysis in team sports 8

1.6 Research significance 11

1.6.1 Theoretical contributions 11

1.6.2 Practical contributions 12

1.7 Limitations 13

1.8 Thesis outline 14

2 Elements of point processes **15**

2.1 Point processes 15

2.1.1 Definition 15

2.1.2 Conditional intensity function 17

2.1.3 Likelihood 17

2.1.4 Compensator 18

2.1.5 Random variate generation 19

2.1.6 Marked point processes 20

2.2 Hawkes processes 22

2.2.1 Definition 22

2.2.2 Random variate generation 23

2.2.3 Immigrant/offspring representation 24

2.2.4 Likelihood calculations for exponential decay 26

2.2.5 Marked Hawkes process 27

3 Proposed model **29**

3.1 Limitations of the marked Hawkes process model 29

3.2 Decoupling the modelling of times and marks 32

3.3 Interpreting the model and its parameters 35

3.4 Parameter estimation via Expectation-Maximisation 36

3.5 Model extensions 38

3.5.1 Covariate dependent conversion rates 38

3.5.2 Event dependent decay rates 39

3.6 Marked spatio-temporal point processes 40

4 Bayesian inference **43**

4.1 Model specification 44

4.1.1 Likelihood 44

4.1.2 Graphical model 45

4.1.3 Prior distributions 46
4.1.4 Impact of prior distributions 47
4.2 Posterior sampling algorithm 48
4.2.1 Gibbs sampling 49
4.2.2 Hamiltonian Monte Carlo 50
4.2.3 HMC implementation in Stan 52
4.3 Sequential updating via importance sampling 55
4.4 Model evaluation 56
4.4.1 Log point-wise predictive density 56
4.4.2 Simulation based validation 57
4.4.3 Performance measures 57

5 Case study: Association football **59**
5.1 Data description 61
5.2 Data exploration 62
5.3 Data pre-processing 65
5.3.1 Cleaning 65
5.3.2 Wrangling 68
5.4 Problem definition 72
5.4.1 Likelihood 73
5.4.2 Training data 73
5.5 Models employed 73
5.5.1 Baseline models 76
5.5.2 Excitation based models 77
5.6 Dealing with model complexity 80
5.6.1 Association rule learning 81
5.6.2 Definition for event sequences 81
5.6.3 Measures of significance 82
5.7 Bayesian inference 84
5.7.1 Gamma process model for the occurrence times . . . 85

5.7.2 Markov chain model for the locations 89
5.7.3 Baseline homogeneous Poisson process model 91
5.7.4 Baseline Markov chain model for the marks 92
5.7.5 Excitation based models for the marks 93
5.8 Model evaluation 97
5.8.1 Log point-wise predictive density 97
5.8.2 Simulation based validation 99
5.9 Parameter description 105
5.9.1 Background mark probability 105
5.9.2 Excitation factor 105
5.9.3 Decay rates 106
5.9.4 Conversion rates 107
5.9.5 Team ability 108
5.10 Recovering hidden structure 112
5.11 Real-time simulation 113

6 Concluding remarks **117**

Bibliography **121**

List of Tables

3.1 Kolmogorov-Smirnov (K-S) test results from the pair-wise comparisons of the distributions of inter-arrival times simulated from a Poisson process, a Hawkes process and a Gamma process with the observed inter-arrival times in football. 32

5.1 A snapshot of the dataset showing a sequence of events with the relevant attributes. 60

5.2 List of teams competing in the 2013/14 season of the English Premier League. 61

5.3 Frequencies of the 22 distinct event types in the dataset. 62

5.4 (Top) Original event sequence with a pair of records for a single CornerAwarded event. (Bottom) Event sequence after removing the outcome = Unsuccessful version of the CornerAwarded event. . . . 66

5.5 (Top) Original event sequence with the erroneous CrossNotClaimed event following a Goal event. (Bottom) Event sequence after swapping the Goal and CrossNotClaimed events. 67

5.6 (Top) Original event sequence with incorrect ordering of Save and SavedShot events. (Bottom) Event sequence after swapping the Save and SavedShot events. 68

5.7 Grouping of event types to actions including the frequency of their observations in the dataset. 69

5.8 Encoding of marks along with their labels and frequencies in the dataset. 70

5.9 Snapshot of the final dataset prepared for modelling. 72

5.10 Zone-wise event frequencies in the training data used for the modelling experiment. 74

5.11 Support $P(x \cap y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y. 83

5.12 $\mathsf{lift}(x \Rightarrow y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y. 83

5.13 Posterior summaries and convergence diagnostics from 3000 posterior samples for selected parameters from the Gamma distribution for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i. 86

5.14 Posterior means of the event specific shape and rate parameters of the Gamma distribution for inter-arrival times in expression (5.3) for in-play (top) and out-of-play events (bottom). The column a_i/b_i gives the posterior mean of the expected value of the Gamma distribution. 88

5.15 Observed transition counts $y_{i\to j}$ from the first 5 states to each zone in the training data. 89

5.16 Posterior means of the multinomial probabilities $\eta_{i\to j}$ for transitions from the first 5 states. 90

5.17 Posterior means of the homogeneous Poisson rates $r_{m,z}$, for the first 5 marks in each zone. 91

5.18 Transition counts $c_{i\to j}$ from the first 5 states to the first 5 marks in the training data. We abbreviate the prefix Home to H in the mark labels. 92

5.19 Posterior means of the multinomial parameters $\theta_{i\to j}$ corresponding to the first 5 states. We abbreviate the prefix Home to H in the mark labels. 93

5.20 Posterior summaries and convergence diagnostics from 1500 posterior samples for selected parameters from the Matrix beta model for the marks in expression (5.8). 94

5.21 Quantifying the impact of prior specifications by computing the ratio of the prior to posterior variance for selected model parameters from the Matrix beta model for the marks in expression (5.8). Ratios much larger than 1 indicate that the prior distributions for the parameters are flat compared to their corresponding posterior distributions. 97

5.22 Cumulative log posterior densities $\widehat{lpd}$ over 10 game periods in the test data for all fitted models along with the number of estimated parameters (N_{par}) in each model. For the Matrix beta models, W is the number of transient events and N is the number of significant event pairs identified in the rule-based framework for reducing model complexity. 98

5.23 Game periods (out of ten) where the model outperforms the baseline in each prediction interval of the validation experiments as designed in Figure 5.13. 103

5.24 Posterior means of the zone dependent background mark probabilities $\delta_{m|z}$ for $z \in \{1, 2, 3\}$ from the Matrix beta model for the marks in expression (5.8). The dots $(\cdot)$ denote $\delta_{m|z}$ values less than 0.01. 106

5.25 Posterior means of team ability parameters ordered by the cumulative ability ($\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$) of the team h to complete a successful pass and retain possession. Team information is incorporated into the event conversion rates using the baseline logit specification in expression (5.9). 109

List of Figures

2.1 A sample path of a simple point process with arrival times t_i are along the x-axis and the counting process $N(t)$ along the y-axis. 16

2.2 An example conditional intensity function $\lambda^*(t)$ for a self-exciting process with background intensity μ. 18

2.3 An illustration of the immigrant/offspring representation of a Hawkes process. Squares indicate immigrants, circles are offsprings, and the crosses denote the occurrence times. 25

2.4 An illustration of the branching structure $\{u_i\}$ from Definition 2.12 for a sequence of events. Squares indicate immigrants, circles are offsprings, and the crosses denote the occurrence times. . . . 26

3.1 Simulated event times from a Poisson process (top) and a Hawkes process (middle) compared against an instance of observed event times from the football dataset (bottom). 30

3.2 Comparing the empirical CDFs of the inter-arrival times of events simulated from a Poisson process (red), a Hawkes process (blue), a Gamma process (green) and observed events in football (purple). ECDFs were computed using 10,000 inter-arrival times in each case. 31

4.1 Graphical model showing conditional dependencies in the Bayesian model specification in Section 4.1. 46

5.1 Visualising the sequence of events leading up to the goal scored by Jack Wilshere for Arsenal against Norwich City, voted as the Goal of the season (2013/14). 63

5.2 Heat map showing the density of ball-touches for Arsenal and Chelsea in their home and away games in the 2013/14 season. In all heat maps the team is attacking to the right, i.e. the opposition goal is to the right. 63

5.3 Heat map showing the density of all shots attempted on goal (left) vs goals (right) across all teams in the 2013/14 season. . . . 64

5.4 Data cleaning workflow showing the steps involved in the three stage process to prepare the dataset for modelling. 65

5.5 Mapping from event location in (x,y) coordinates to zones. 71

5.6 Trace plot of the posterior samples for selected parameters across multiple chains from the Gamma process model for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i. 86

5.7 Pair-wise correlations with marginals along the diagonal for selected model parameters from the Gamma distribution for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i. 87

5.8 Trace plot of the posterior samples for selected parameters across multiple chains from the Matrix beta model for the marks in expression (5.8). 94

5.9 Pair-wise correlations with marginals along the diagonal for selected model parameters from the Matrix beta model for the marks in expression (5.8). 95

5.10 Visualising the impact of prior specifications by overlaying the posterior and prior densities for selected model parameters from the Matrix beta model for the marks in expression (5.8). 96

5.11 Proportion of importance sampling weights greater than a series of progressively increasing thresholds where $R = 500$ is number of posterior samples. The weights are calculated for the first game period in the test data given the event history up to time T in minutes. 99

5.12 Posterior distributions for $\beta_{1\to 3|1}$ before and after updating given the event history up to time T of the first game period in the test data. 100

5.13 Design of the validation experiments, where the models are evaluated in four separate two-minute prediction intervals within each game period in the test data, given the observed history of events before each interval. 101

5.14 Predictive distributions of the number of successful passes by the home team (Arsenal) in four intervals of the first half in the game between Arsenal and Tottenham Hotspur in the test data. The observed count (truth) is given by the vertical dashed line. 102

5.15 Scoring rules for selected event types within a randomly chosen game in the test set with the prediction interval intervals along the x-axis. 104

5.16 Posterior means of the event conversion probabilities $\gamma_{m_j\to m_i|z_i}$ for a selection of event pairs corresponding to the location $z = 2$ from the Matrix beta model for the marks in expression (5.8). The $\gamma_{m_j\to m_i|z_i}$'s are computed for a hypothetical match-up where the baseline team West Ham United is chosen as both the home as well as the away team to negate the impact of team abilities. . . . 107

5.17 Posterior distribution of the parameters $\omega_{h,\text{Home\_Pass\_S}}$ in (a) and $\omega_{h,\text{Away\_Pass\_S}}$ in (b), from the baseline logit specification for incorporating team abilities in expression (5.9). Teams are ranked in the decreasing order of the means of their respective posterior distributions shown by the overlayed vertical lines. 110

5.18 (a) Posterior distribution of $\omega_{h,\text{Home\_Shot}} + \omega_{h,\text{Away\_Shot}}$, the cumulative ability of a team h, relative to West Ham (baseline), to attempt a shot on goal. (b) The number of shots, passes completed in the attacking third and shots per pass completed in the attacking third (S/P) for each team in the training data. 111

5.19 (a) Team rankings based on the cumulative ability to trigger a particular event type. For example, the column Pass, ranks teams in the decreasing order of their respective posterior means of $\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$. (b) The final positions of the teams in the league table of the 2013/14 season taken from www.whoscored.com. The team rankings estimated using the cumulative passing ability in (a) is the best predictor of the final positions in the league table in (b). 112

5.20 Branching structure probabilities for events in the first 4 minutes of the game between Chelsea and Hull City in the 2013/14 season of the English Premier League. The highlighted event Home_Shot has a higher probability of being an offspring of the event Home_Out_Corner than being an offspring of the more recent Home_Pass_S event. 114

5.21 Forecasting the probability of observing at least one Home Shot event in 1-minute intervals over the first half of the game between Arsenal and Tottenham Hotspur in the test data. Intervals with observed Home Shot events are highlighted using dotted lines. . 115

Acknowledgements

I am extremely grateful to my supervisors, Dr Ioannis Kosmidis and Professor Petros Dellaportas, for their dedicated support and constant guidance. I am indebted to fellow research students David Selby, for his valuable tips on R programming, and Tim Stumpf-Fétizon for his feedback and stimulating discussions related to the project. Thanks also go to Stratagem Technologies for providing football data in a convenient format.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. The work contained within is original and has not been submitted previously for a degree at any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made.

Abstract

We consider the modelling of a collection of marked point processes where the occurrence rate depends on past occurrences within the process. Building on a traditional model for point processes, the Hawkes process, we restrict its characteristic properties exclusively to the space of marks, providing the freedom to specify a different model for the occurrence times. The main idea is to use the decomposition of a multivariate density function to decouple the joint modelling of the event types (marks) and the occurrence times. We develop a Bayesian framework for the inference and prediction of these flexible marked point processes that can be applied to a wide range of applications. We present a case study on the modelling of event-sequences from association football, where the sequence of game events can be treated as a marked spatio-temporal point process. We provide inferences about previously unquantified measures governing the dynamics of the game as well as predict the occurrence of events of interest, such as goals, corners or fouls, in a specified interval of time.

Nomenclature

$\emptyset$	Empty set
$\mathbb{E}$	Expectation
$\mathbb{P}$	Probability
$\mathbb{N}$	set of Natural numbers
$\mathbb{R}$	set of Real numbers
$\mathbb{R}^+$	set of Real numbers greater than or equal to 0
$\mathbb{Z}$	set of Integers

Chapter 1

Introduction

1.1 Motivation

Will my football team score a goal in the next 10 minutes?
When will the next major earthquake strike Indonesia?
How long do we have before AI takes over the world?

We often find ourselves facing the task of predicting when a particular event might happen. Some events might be harder to predict than others, nevertheless, arriving at the best possible predictions to many such questions is extremely useful.

Phenomena that are observed as a sequence of events happening over time can be represented using point processes. While point processes can be used to describe a random collection of points in any general space, we limit ourselves to the case in which the points denote events that occur along a time axis. Such point processes, having a natural order in which the points occur, are suitable for a wide range of real-world applications and are well studied in probability theory (see, for example, Daley and Vere-Jones, 2003).

Processes in which points are identified only by the occurrence times are referred to as univariate point processes. Multivariate point processes, on the other hand, are those in which two or more types of points are observed.

For example, in a queuing system, the arrivals and departures of customers would be two types of points that are observed.

Multivariate point processes are specified by associating a random variable, say m, to each point in a univariate point process, where the realised value of m gives the point type. If m is allowed to be a general random variable, not restricted to be a category giving the point type, then we refer to m as a mark and the process as a marked point process. An example of a marked point process with continuous marks is in seismology, where the magnitude of an earthquake is recorded in addition to the time of occurrence.

When event sequence data are analysed using point process models, an important distinction is between *empirical* models and *mechanistic* models as noted by Diggle (2013). Empirical models have the solitary aim of describing the patterns in the observed data, while mechanistic models go beyond that and attempt to capture the underlying scientific process that generated the data. Mechanistic models for marked point processes are typically specified using a joint conditional intensity for the occurrence times and the marks and in general are not flexible enough to be applied to real-world datasets. The joint modelling of the components of the process can also be challenging and it is common to make strong restrictive assumptions like separability (González et al., 2016) to simplify the model. The primary focus of this research is to develop a flexible mechanistic modelling framework for marked point processes that are suitable for a wide range of applications.

The focus area of this research is motivated by the problem of modelling event sequences from association football, with the aim of quantifying the underlying dynamics of the game. Football is one of the most popular team sports and is an example of an invasive sport, where two opposing teams compete for the possession of the ball with the dual objective of attacking to score a goal and defending against attacks by the opposition. Over the

last decade, there has been a concerted effort to record events that happen during the course of games at high frequency and accuracy. The resulting data are directly relevant in the development of game strategies, team and player performance evaluation and in enhancing the viewing experience of televised games.

Most analyses in football are typically done manually by watching video footage or using simple frequency analysis of match events. Hence, there is a huge scope to improve the efficiency of the data-analytic methods as well as the quality of performance evaluation. However, the analysis of football data is mathematically challenging due to the continuous interaction between players within and across the two teams. We recognised that marked point processes are well suited to analyse football event data and provide an excellent foundation to achieve our goal of describing the game dynamics.

1.2 Problem statement and research goals

Traditional models for marked point processes are typically specified using a joint conditional intensity function for the occurrence times and the marks. A joint specification can prove to be quite restrictive and inconvenient in many cases like the modelling of event sequences observed in football.

Problem Statement

We wish to (a) restrict the characteristic properties of a marked point process model exclusively, say, to the marks and (b) have the freedom to specify a different model for the occurrence times. How do we develop a general modelling framework for the prediction and inference from such a flexible marked point process?

Essentially, we tackle this problem by building on the decomposition of a multivariate density function in Cox (1975, Expression 2).

During the course of developing a flexible modelling framework for marked point processes, we set ourselves the following research goals. We broadly divided the goals into those aligned towards developing methodology and those related to the application.

Methodology-specific goals

1. **Simulation**: Develop a framework to simulate the marked point process in the interval $(T, T + d)$ where T is the current time and $d > 0$ represents the time duration of prediction.
2. **Capture dependencies**: Properly account for dependencies on past occurrences within the process and process-specific characteristics.
3. **On-line inference**: Develop machinery to run simulations efficiently and make predictions in real time, after updating the parameter estimates based on the newly observed data.
4. **Validation**: Develop a validation framework for evaluating the predictive quality of the fitted models.

We present a case study on the modelling of events sequences from association football, where we build a game simulator that can simulate the entire sequence of events (times and event types) from the start or any intermediate point till the end of the game.

Application-specific goals

4. **Predictions**: In real time, predict (a) game outcome probabilities and (b) team-specific probabilities of any event, e.g. goal scored, in the next d minutes.
5. **Parameter descriptions**: Develop parameter interpretations and describe how they quantify and provide insight into the dynamics of the

game.

6. **Impact of covariates**: Quantify the impact of covariates, such as the team information, by incorporating them into the model.

1.3 Theoretical foundation

The most typical point process is the Poisson point process (see, for example, Kingman, 1993), named from the fact that the number of points observed in an interval of the process is a random variable with a Poisson distribution. The Poisson process is memory-less, in the sense that the probability of a point occurring in any interval is independent of the past occurrence times. Due to its convenient mathematical properties, it is widely used to model random event occurrences in time, for example, in queuing theory (Kleinrock, 1975) to model the arrival of customers to a shop or phone calls at an exchange.

The memory-less property of the Poisson point process is unsuitable for many real-world phenomena, which led to the development of models in which the occurrence rate depends on past occurrences within the process. Hawkes processes, a mathematical model for *self-exciting* processes, was proposed in Hawkes (1971). Like any temporal point process, it can be used to model a sequence of arrivals of some type over time, for example, earthquakes in Ogata (1998). Each arrival excites the process in the sense that the chance of a subsequent arrival is increased for some period of time after the initial arrival. The excitations from previous arrivals add together and as such, it is a non-Markovian extension of the Poisson process.

Marked Hawkes processes are typically specified using a joint conditional intensity function for the occurrence times and the marks (see, for example, Rasmussen, 2013, Expression 2.2). The marked Hawkes process model captures the magnitudes of all cross-excitations between the various event

types as well as the rate at which these excitations decay over time. Excitation leads to clustering of events in time as the process is driven by an intensity that increases with every arrival for a short period of time. However, in applications like the event sequences observed in football, the events tend not to cluster in time and the marked Hawkes process model is not suitable. The joint modelling of the times and the marks has to be decoupled to restrict the excitation property of the process exclusively to the marks.

We take advantage of the decomposition of a multivariate density function that motivated the partial likelihood in Cox (1975). The joint conditional distribution for a marked point process can be factorised into a probability density function for the next event time conditioned on the past occurrences and a probability distribution function for the event mark conditioned on the time of occurrence and the past. Therefore, an alternate approach to specify a marked point process model would be to specify the conditional distribution functions for the times and the marks separately. We derive the conditional distribution function for the marks in the case of a marked Hawkes process from its typical joint conditional intensity specification. Crucially, we then have the freedom to specify a completely new density function for the times best suited to our application. As a result, we construct a marked point process model that retains the characteristic properties, like excitation in Hawkes processes, in the model for the marks while avoiding the clustering of event times.

1.4 Research methodology

The methodologies developed in this project are focused towards the research goals set in Section 1.2. We develop a modelling framework for flexible marked point processes by decoupling the joint modelling of the occurrence times and the marks. Specifically, we carry out the derivations

of the probability distribution function for marks in the case of the marked Hawkes process model, a choice purely driven by its suitability for the application of event sequences observed in football.

We find the excitation framework of the marked Hawkes process model appropriate for the event sequences in football, as any event in the sequence is likely to be triggered by one or more of the previous events. For example, following a corner kick, the next event is almost surely one among a shot on goal, a defensive clearance or a claim by the keeper. In that sense, the corner kick excites the occurrence chance of those three event types in the immediate future. The proposed model, based on the excitation framework of marked Hawkes processes, captures the magnitudes all cross-excitations between the various event types as well as the rate at which these excitations decay over time. The model also allows the incorporation of covariates such as team information in a direct way to capture the relative abilities of teams.

We provide details on the parameter estimation for such flexible marked point processes via an EM (Expectation-Maximisation) algorithm (Dempster et al., 1977). In addition, we develop a more detailed Bayesian approach, keeping an eye on our goal of online inference. The Bayesian paradigm of updating one's beliefs based on new information is well suited to such a task.

Most analyses in this project have been carried out using the software R by R Core Team (2019). The statistical modelling platform Stan by Stan Development Team (2020) is used for performing Bayesian inference via a variant of the Hamiltonian Monte Carlo algorithm, originally proposed by Duane et al. (1987), to generate samples from the posterior distribution of the parameters.

After obtaining samples of the model parameters using training data, we develop an algorithm to update the parameter samples based on test data

using importance sampling (Kahn and Harris, 1951). The idea is to resample with replacement from the posterior samples using unequal weights that are proportional to the ratio of the likelihood of the complete data (train and test) to the likelihood of the training data only. We also develop a simulation framework, where we detail the steps to simulate the specified marked point process up to any time in the future given the history of the process. The parameter updating followed by the process simulation is implemented efficiently to make real-time predictions of game outcomes and event-specific occurrence probabilities.

1.5 Review of event sequence analysis in team sports

Over the last decade, the availability of spatio-temporal data from team sports has inspired research into the application of statistical methods for team and player performance evaluation. A comprehensive survey of the recent research efforts in spatio-temporal analysis of team sports is provided in Gudmundsson and Horton (2017). There are two primary types of spatio-temporal data collected from team sports. *Movement data* consists of samples of timestamped locations in the plane tracking the movement of all players and the ball during the game. Player movement is captured using fixed cameras in optical tracking systems, that process the images to compute the trajectories. *Event data* streams, on the other hand, record the sequence of events that occur during the game. Event data is manually collected by trained analysts who watch video feeds of the games through a special annotation software. Companies like Opta provide data of both the movement and event formats.

As our work is motivated by the availability of event data from football, we limit ourselves to the review of research using event data streams. Event data are less dense than movement data, but richer in the sense that it

contains more information about what is happening in the game. Events broadly fall into two categories; player events such as passes and shots; and stoppage events such as fouls, end of game etc. Every event is annotated with, among others, a timestamp, location (i.e., a (x,y) position), an event type (e.g., pass, foul) and the players involved.

A popular research area using event data is the network analysis of player interaction. Models for player interaction can quantify a team's playing style as well as the importance of an individual player within the team. Players are identified as nodes of the network and are connected using directed edges whose weights are proportional to the number of successful passes between the two players. Passing networks were first applied to team sports in Passos et al. (2011) to study a team's collective behaviour in water polo. Grund (2012) studied the degree centrality of passing networks in football, which quantifies the importance of nodes in the network based on the number of edges. They showed that teams that rely heavily on key players performed relatively worse. Duch et al. (2010) used flow centrality to assess player performance by capturing the fraction of times that a player intervenes in those paths that result in a shot on goal. They also take into account defensive behaviour by letting each player start a number of paths proportional to the number of balls they recover. Clemente et al. (2015) studied the density and heterogeneity of passing networks and showed how high heterogeneity leads to formation of sub-communities, meaning there is a low level of cooperation between the players of a team. Pena and Touchette (2012) looked at other centrality measures such as closeness and eigenvector centrality as well as clustering in football passing networks.

Another use of event data is in the identification of *plays*, i.e., sequences of passes between a small group of players that occurs repeatedly. In Borrie et al. (2002), passes are identified by the zones in the pitch they start and end in and frequently occurring sequences are detected by also taking into

account the time intervals between passes. Wang et al. (2015) proposed an unsupervised approach to automatically detect tactical patterns in football. They present the Team Tactic Topic model based on Latent Dirichlet Allocation to identify tactics from pass sequences. Interesting visualizations are provided for the most successful tactics as well as how a team's tactical patterns evolve over a season. Van Haaren et al. (2016) also look at automatic discovery of patterns in attacking strategy. They use a data-driven approach to determine a number of spatial features about the areas occupied during a continuous possession phase of a team. The features are then used to cluster similar phases together to identify frequently occurring event sequences within the cluster. Decroos et al. (2017) divide the game using overlapping windows to create subsequences of events to use as a feature to predict a goal event in the near future. They compute similarity between subsequences using Dynamic Time Warping, a distance measure for time-dependent sequences.

Extracting game states from event sequences to quantify the value of player actions or to make predictions of the game outcome is another interesting area of research. Routley and Schulte (2015) used Markov decision processes for valuing player actions in Ice Hockey. Game states are derived from contextual features like game score and time remaining along with the recent history of events. The associated reward for an action in the Markov decision process gives the value of the player action. A similar approach based on game states is taken in Decroos et al. (2018) to value player actions in football. They train a classification model to calculate the probability a game state will to lead to a goal in the near future, where each game state is described using over 150 features. The value of a player action is then calculated by the shift in the predicted goal probability before and after the action. Other approaches for predicting goal probabilities based on a current game state are by Mackay (2017) and Robberechts et al. (2019). Approaches based

on game states involve significant effort into feature engineering and with the use of learning algorithms like gradient boosting that limit parameter interpretations, the methods provide little, if any, insight into the dynamics of the game.

1.6 Research significance

We discuss the significance of this research towards advancing the theory of marked point processes as well as team sports like football.

1.6.1 Theoretical contributions

The idea of using the decomposition of a multivariate density function to restrict the characteristic property of a process exclusively, for example, to the space of marks, is novel and provides us the freedom to specify a different model for the occurrence times best suited to the application.

Although the derivation of the model specification has been carried out for marked Hawkes processes, the modelling framework is suitable for any marked point process model that is specified using a joint conditional intensity function for the times and marks. We found the marked Hawkes process model to be powerful and yet elegant in the way it can capture event interactions over arbitrary lengths of time, which have made them popular for many real-world applications.

The Bayesian modelling framework developed in this thesis can be readily applied to other applications, especially when on-line inference is necessary. We discuss a formal approach to evaluate the goodness of fit of point process models using the out-of-sample log predictive density. We implement a framework for updating parameters followed by process simulation, that can be used efficiently to make predictions in real-time. We also adapt an existing method for estimation of the traditional Hawkes model (Veen and

Schoenberg, 2008), to develop an EM algorithm for parameter estimation of our decoupled model. We show how the decoupled model preserves the immigration-offspring representation of the process and how the associated branching structure can be used in the estimation procedure.

1.6.2 Practical contributions

The major focus of existing methods in team sport analysis appear to be tailored towards individual player performance evaluation or identifying specific patterns in team play. And as discussed in Section 1.5, most approaches take the route of summarising the event data into compact representations like networks and game states. However, in this project, we take a more holistic approach to study football as a dynamical system and model the entire sequence of events within a game. Such a model, that captures all event interactions, is well suited to predict the occurrence of the rare goal scored events, that determine the outcome of the game.

To the best of our knowledge, the use of point processes to model the event data from team sports is novel. In fact, we have not come across any other attempt to model the entire sequence of events within a game. Using the point process model we have developed, we are able to create a game simulator that simulates the entire sequence of events (times and event types) from the start or any intermediate point till the end of the game. The simulations are fast enough to be run in real-time and we can obtain instantaneous predictions of goal probabilities, game outcomes or other quantities of interest such as possession ratio or passing accuracy etc. We believe these predictions would enhance, among others, the viewing experience of televised games.

The parameters of the model also provide us with valuable insight into the dynamics of the game. The excitation framework of the proposed model captures both the magnitudes and durations of all pairwise event interac-

tions. The model along with its parameters can be used to develop a deeper understanding of the game-play by the coaching staff and inform strategic decision making. The modelling framework and statistical methodology developed in this project can be readily applied to many other team sports like rugby, hockey, basketball etc. As none of the methods have been tailored specifically to football or even sports for that matter, they can also be applied to a wide range of applications that generate event data streams.

1.7 Limitations

The following are the two key limitations of our research which we identify as avenues for future work in the modelling of marked point processes.

- **Linear excitation:** The traditional Hawkes process model is a linear self-exciting process. Self-exciting in the sense that an arrival increases the intensity of the point process, and linear in the sense that the excitations from different arrivals add up. These two assumptions can be quite restrictive, however moving away from either of them would invalidate many existing theoretical results because the immigrant-offspring representation is no longer viable. Non-linear generalisations of the Hawkes process have been considered in Brémaud and Massoulié (1996), and processes allowing inhibition along with excitation in Mei and Eisner (2017).

- **Exchangeability of games:** In our modelling framework, we do not account for the evolution of the parameters over games and in that sense the natural order in which the games take place is ignored. Having a vector auto-regressive structure (Rue and Held, 2005) to model the time-varying parameters over games could be a valuable addition to the hierarchical Bayesian framework we have developed.

1.8 Thesis outline

The thesis is structured as follows:

- Chapter 2 presents the essential elements of point processes in general before taking a deep dive into Hawkes processes, a classical model which is extended to tackle the application in football.
- Chapter 3 discusses the limitations of Hawkes processes and the need to decouple the joint modelling of the marks and occurrence times. A flexible modelling framework is proposed and an EM algorithm for its parameter estimation is developed. The key extensions to the model are also discussed.
- Chapter 4 presents a Bayesian framework for the inference and prediction of marked point processes. The Hamiltonian Monte Carlo algorithm for sampling from the posterior distribution and its software implementation are discussed. The methods for model evaluation including the set-up of a simulation framework and the performance measures used for its validation are also detailed.
- Chapter 5 starts by exploring the football event data and the different kinds of analyses such data can be used for. Details on the preprocessing steps are provided before defining the modelling task. The baseline and excitation based models that are employed in the case study are then specified. An association rule based approach to reduce model complexity is developed, before discussing the Bayesian inference for all the fitted models. The results from two approaches for evaluating the accuracy of the models are then presented. Finally, a complete description of all the estimated parameters are provided and the insight they provide into football are discussed.
- Chapter 6 provides some concluding remarks.

Chapter 2

Elements of point processes

Sequences of events over time are conveniently represented using point processes making them suitable for a wide range of real-world applications. In this chapter, we discuss the essential elements of point processes in general before taking a deep dive into Hawkes processes, a classical model which we will build on to tackle the application we are interested in.

2.1 Point processes

First, we introduce point processes and define the key concepts of filtration and the conditional intensity function, thereby setting essential notation. We then discuss the point process likelihood, the compensator function, and the random time change theorem, which are useful for parameter estimation and random variate generation. Finally, we define marked point processes, its key properties and discuss the separability of the conditional intensity function.

2.1.1 Definition

A point process is a model for a sequence of arrivals into a system and is defined as follows.

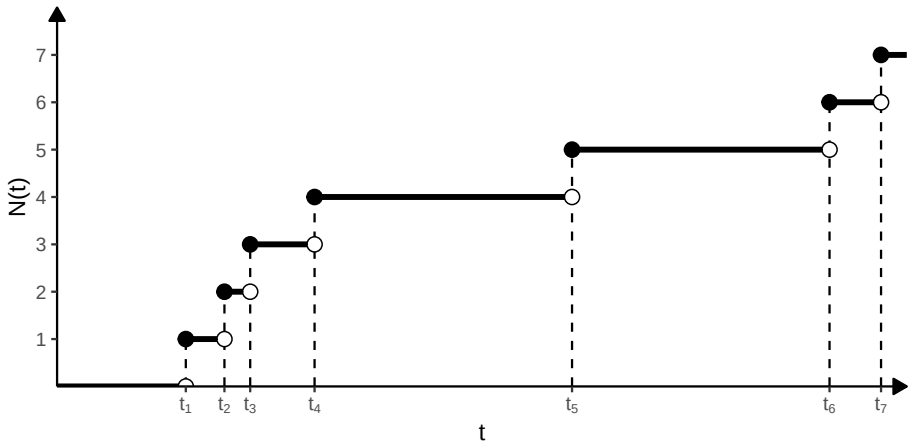


Figure 2.1: A sample path of a simple point process with arrival times t_i are along the x-axis and the counting process $N(t)$ along the y-axis.

Definition 2.1 *Let* $\mathbf{x} = \{t_i\}$ *be a sequence of ordered points such that* $\forall\ i \in \mathbb{N}$, $t_1 \geq 0$ *and* $t_i < t_{i+1}$. *Then,* $\mathbf{x}$ *is a sample path of a simple point process on* $\mathbb{R}^+$.

The point process in Definition 2.1 is simple in the sense that occurrences cannot be simultaneous. Any point process has an associated counting measure $\nu(I)$, for any Borel subset I of $\mathbb{R}^+$, defined as the number of occurrences in the set I.

$$\nu(I) = \#\{i : t_i \in I\}\,.$$

Point processes can be specified using its counting measure, for example, $\mathbf{x}$ is a sample path of a Poisson process, if and only if (Daley and Vere-Jones, 2003, Theorem 2.3.I), for all sets I that can be represented as the union of a finite number of intervals of finite length, $\mathbb{P}\{\nu(I) = 0\} = \exp(-\lambda\ell(I))$, where $\lambda > 0$ and $\ell(\cdot)$ denotes the Lebesgue measure.

Figure 2.1 is a sample path of a simple point process, where the arrival times are along the x-axis and the counting process $N(t)$ along the y-axis, which is the step function

$$N(t) = \nu([0,t]) \qquad (0 < t < \infty)\,,$$

that counts the number of occurrences up to time t.

Definition 2.2 *We define the history or filtration $\mathcal{F}_t$ at time t of the point process as $\mathcal{F}_t = \{t_j : t_j \in \mathbf{x} \text{ and } t_j \leq t\}$.*

Henceforth, we shall work under the setting where we observe a process from its beginning at time $t = 0$ and $\mathcal{F}_0 = \emptyset$.

2.1.2 Conditional intensity function

Point processes are typically characterised using the conditional intensity function defined in Definition 2.3 adopted from Laub et al. (2015, Definition 3). Indeed, if the conditional intensity function exists it uniquely characterises the point process (Daley and Vere-Jones, 2003, Proposition 7.2.IV).

Definition 2.3 *The conditional intensity function, given a counting process $N(t)$ with filtration $\mathcal{F}_t$, is defined as*

$$\lambda^*(t) = \lambda(t|\mathcal{F}_t) = \lim_{h \to 0} \frac{\mathbb{E}[N(t+h) - N(t)|\mathcal{F}_t]}{h}.$$

We abbreviate $\lambda(t|\mathcal{F}_t)$ to $\lambda^*(t)$ to avoid specifying the filtration $\mathcal{F}$ explicitly. Self-exciting processes are those in which an arrival causes the conditional intensity function to increase causing a clustering of arrival times. Figure 2.2 is an example illustration of the conditional intensity function of a self-exciting point process.

2.1.3 Likelihood

Assume we have observed a point process $\mathbf{x} = \{t_1, t_2, \ldots, t_n\}$ on $[0, T)$ for some fixed time $T > 0$, and no points have occurred before 0. Then, by Daley and Vere-Jones (2003, Proposition 7.2.III), the likelihood function is

$$\left[\prod_{i=1}^{n} \lambda^*(t_i)\right] \exp\left\{-\int_0^T \lambda^*(u)\,\mathrm{d}u\right\}.$$

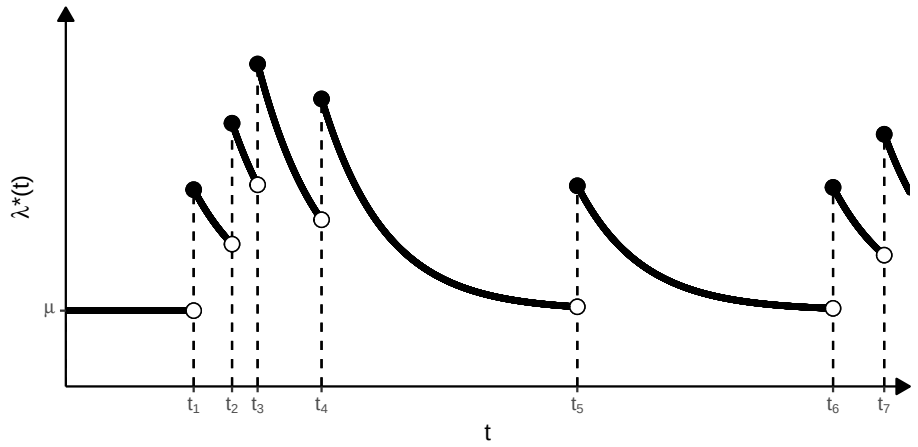


Figure 2.2: An example conditional intensity function $\lambda^*(t)$ for a self-exciting process with background intensity μ.

And the log-likelihood for the interval $[0, T)$ is

$$\sum_{i=1}^{n} \log\left(\lambda^*(t_i)\right) - \int_0^T \lambda^*(u)\,\mathrm{d}u\,. \tag{2.1}$$

2.1.4 Compensator

The integral of the conditional intensity function

$$\Lambda^*(t) = \int_0^t \lambda^*(s)\,\mathrm{d}s\,,$$

is known as the compensator of the point process. The compensator is a key quantity because of the mapping property of point processes described in Kingman (1993, Mapping Theorem). If the state space is mapped into another space, the transformed random points again form a point process. Specifically, when the compensator is used as the mapping function, the resulting process is a Poisson process with unit rate. For point process over time, this is referred to as the random time change theorem stated in Theorem 2.4.

Theorem 2.4 *(Daley and Vere-Jones, 2003, Theorem 7.4.I) Let $N(t)$ be a counting process with a strictly positive conditional intensity function $\lambda^*(t)$ and compensator*

$\Lambda^*(t)$ that is a.s. bounded, then under the random time change $t \to \Lambda^*(t)$, the transformed counting process $\tilde{N}(t) = N(\Lambda^{*-1}(t))$ is a Poisson process with unit rate.

2.1.5 Random variate generation

We discuss two fundamental approaches to simulating sample paths from point processes.

Inversion

The basic idea of the inversion method is to simulate a unit rate Poisson process (this is just a cumulative sum of a series of independent exponential random variables with rate one) and transform these into the desired point process using the inverse compensator function (Daley and Vere-Jones, 2003, Algorithm 7.4.III). The inversion method is a direct consequence of the random time change theorem for point processes and is stated in Proposition 2.5, adopted from Daley and Vere-Jones (2003, Theorem 7.4.I).

Proposition 2.5 *If $\{u_i\}_{i\in\mathbb{N}}$ is a sample path of a unit rate Poisson process on $\mathbb{R}$, and $t_i = \Lambda^{*-1}(u_i)$, then $\{t_i\}_{i\in\mathbb{N}}$ is a sample path of a point process with intensity $\lambda^*(t_i)$.*

Thinning

A Poisson point process is said to be homogeneous if it has constant intensity. The standard way to generate an in-homogeneous Poisson point process driven by intensity function $\lambda(\cdot)$ is via thinning (Lewis and Shedler, 1979). The intuition is to generate a 'faster' homogeneous Poisson point process, and remove points probabilistically so that the remaining points satisfy the time-varying intensity $\lambda(\cdot)$. It is required that the homogeneous process' rate M cannot be less than $\lambda(\cdot)$ over $[0, T]$. Formally the process is described by Algorithm 1, adopted from Laub et al. (2015, Algorithm 1).

Algorithm 1 Generate a Poisson point process by thinning.

procedure Thinning($T, \lambda(\cdot)$, M)
  require: $\lambda(\cdot) < M$ on $[0, T]$
  $P \leftarrow [], t \leftarrow 0$
  while $t < T$ **do**
    Generate next candidate point:
    $E \leftarrow Exp(M), t \leftarrow t + E$
    Keep it with some probability:
    $U \leftarrow Unif(0, M)$
    if $t < T$ and $U \leq \lambda(t)$ **then**
      $P \leftarrow [P, t]$
    end if
  end while
  return P
end procedure

2.1.6 Marked point processes

The events constituting a point process can carry additional information (e.g. event type), generally referred to as marks. For any such marked point process, the times $\{t_i\}$ at which the events occur constitute a process by itself and is called the *ground process* denoted by N_g. We can now formally define the marked point process and its main properties, adopted from Daley and Vere-Jones (2003, Definitions 6.4.I and 6.4.III).

Definition 2.6 *A marked point process (MPP), with times in the completely separable metric space (c.s.m.s.) $\mathcal{T}$ and marks in the c.s.m.s $\mathcal{M}$, is a point process $\{(t_i, m_i)\}$ on $\mathcal{T} \times \mathcal{M}$ with the additional property that the ground process $\{t_i\}$ is itself a point process on $\mathcal{T}$.*

A completely separable metric space means that there exists a sequence $\{c_n\}_{n=1}^{\infty}$ of elements of the space such that every non-empty open subset of the space contains at least one element of the sequence.

Definition 2.7 *A multivariate point process is a marked point process where marks take values from the finite set $\{1, \ldots, M\}$ for some finite integer M.*

The next two definitions characterise the two important types of independence relating to the mark structure of MPPs.

Definition 2.8 *An MPP has independent marks if, given the occurrence times* $\{t_i\}$*, the marks* $\{m_i\}$ *are mutually independent random variables such that the distribution of the mark* m_i *depends only on the corresponding time* t_i*.*

Definition 2.9 *An MPP has unpredictable marks if the distribution of the mark* m_i *at time* t_i *is independent of the past history of times and marks* $\{(t_j, m_j) : j < i\}$*.*

Conditional intensity function

A marked point process is typically specified using its joint conditional intensity function

$$\lambda^*(t, m) = \lambda_g^*(t) f^*(m \mid t)\,, \tag{2.2}$$

where $\lambda_g^*(t)$ is the conditional intensity of the ground process N_g and $f^*(m \mid t)$ is the conditional density of the mark at time t. Both $\lambda_g^*(t)$ and $f^*(m \mid t)$ are conditioned on $\mathcal{F}_{t^-}$, the filtration of the marked point process up to (but not including) t. If $f^*(m \mid t)$ is a proper density function over the mark space that integrates to 1, we have

$$\lambda_g^*(t) = \int_{\mathcal{M}} \lambda^*(t, m)\, \mathrm{d}m\,. \tag{2.3}$$

Separability

Separability of the conditional intensity function for a marked point process assumes that the conditional intensity has the following simpler form (see, for example, González et al., 2016, Section 6.5),

$$\lambda^*(t, m) = \lambda_g^*(t) f^*(m)\,. \tag{2.4}$$

Separability is a rather restrictive assumption that implies that the conditional distribution of the mark is independent of the time of occurrence t.

However, if we are able to assume separability, it is convenient since the sequence of marks can then be modelled separately from the sequence of times.

Likelihood

By Daley and Vere-Jones (2003, Proposition 7.3.III), the log-likelihood for a marked point process with mark dependent conditional intensity $\lambda^*(t, m)$ and n observed events, is

$$\sum_{i=1}^{n} \log\left(\lambda^*(t_i, m_i)\right) - \int_0^T \int_{\mathcal{M}} \lambda^*(v, m)\, \mathrm{d}m\, \mathrm{d}v\,. \tag{2.5}$$

Substituting expressions (2.2) and (2.3) in (2.5) gives

$$\sum_{i=1}^{n} \log\left(\lambda_g^*(t_i)\right) + \sum_{i=1}^{n} \log\left(f^*(m_i \mid t_i)\right) - \int_0^T \lambda_g^*(v)\, \mathrm{d}v\,. \tag{2.6}$$

2.2 Hawkes processes

With the essential background and core concepts detailed in Section 2.1, we now turn to discussing the Hawkes process. Hawkes processes (HP) are point processes whose defining characteristic is that they *self-excite*, meaning that each arrival increases the rate of future arrivals for some period of time.

2.2.1 Definition

Definition 2.10 *(Hawkes, 1971, self-exciting point process) Consider a counting process $N(\cdot)$ with associated filtration $\mathcal{F}_t$ and conditional intensity function of the form*

$$\lambda^*(t) = \mu + \int_{-\infty}^{t} g(t-u) dN(u)\,,$$

where $\mu > 0$ and $g(u) \geq 0$ for $u \geq 0$. Such a process $N(\cdot)$ is called a Hawkes process.

The parameter μ is referred to as the background intensity of the process and the function $g(\cdot)$ expresses the positive influence of the past events on the current value of the intensity process. Hawkes (1971) also showed that if we assume stationarity we have

$$0 < \int_0^\infty g(u)\,\mathrm{d}u < 1\,. \tag{2.7}$$

We adapt the Definition 2.10 to a setting where we observe a sequence of non-negative arrival times into a system, where $\mathcal{F}_0 = \emptyset$. In order to provide a more intuitive interpretation of the Hawkes intensity function, let us use $\{t_1, t_2, \ldots, t_k\}$ to denote the observed sequence of arrival times up to time t and rewrite the Hawkes conditional intensity as

$$\lambda^*(t) = \mu + \sum_{t_i < t} g(t - t_i)\,. \tag{2.8}$$

The Hawkes intensity is therefore the combined effect of the background intensity and the sum of all excitations caused by the past events.

Remark 2.11 *Definition 2.10 describes a Hawkes process that is linear in the sense that the excitations from different arrivals add up. Unless otherwise specified, the HPs in this thesis will refer to this linear form.*

2.2.2 Random variate generation

A standard approach to simulate a Hawkes process is using the modified thinning algorithm described in Ogata (1981). It is common for the intensity to be non-increasing in periods without any arrivals. This implies that for $t \in (t_i, t_{i+1}]$, $\lambda^*(t) \leq \lambda^*(t_i)$. So the rate M of the 'faster' homogeneous Poisson process can be updated during each simulation. Algorithm 2, adopted from Laub et al. (2015, Algorithm 2) outlines the procedure.

Algorithm 2 Generate a Hawkes process by thinning.

procedure HawkesByThinning($T, \lambda^*(\cdot)$, M)
  require: $\lambda^*(\cdot)$ is non-increasing in periods on no arrivals
  $\varepsilon \leftarrow 10^{-10}$ (some small positive value)
  $P \leftarrow [], t \leftarrow 0$
  while $t < T$ **do**
    Find new upper bound:
    $M \leftarrow \lambda^*(t + \varepsilon)$
    Generate next candidate point:
    $E \leftarrow Exp(M), t \leftarrow t + E$
    Keep it with some probability:
    $U \leftarrow Unif(0, M)$
    if $t < T$ and $U \leq \lambda^*(t)$ **then**
      $P \leftarrow [P, t]$
    end if
  end while
  return P
end procedure

2.2.3 Immigrant/offspring representation

The standard definition of the Hawkes process does not directly provide any intuition towards the causality of the event occurrences. Any event is triggered from an intensity contributed to by all previous events and the background intensity as specified in expression (2.8). The only exception is for the first event in the sequence that is triggered solely from the background intensity. However, we may want to assume a causal constraint that any event is triggered by exactly one of the previous events or the background and as this triggering is unobserved, we wish to recover this hidden branching structure in addition to modelling the intensity of the point process.

Hawkes and Oakes (1974) showed that the intensity function specified in expression (2.8) indicates that the Hawkes process is a generalised branching Poisson process (Lewis, 1969) and called it the Poisson cluster process. Hawkes and Oakes (1974) presented a probabilistic model of events in con-

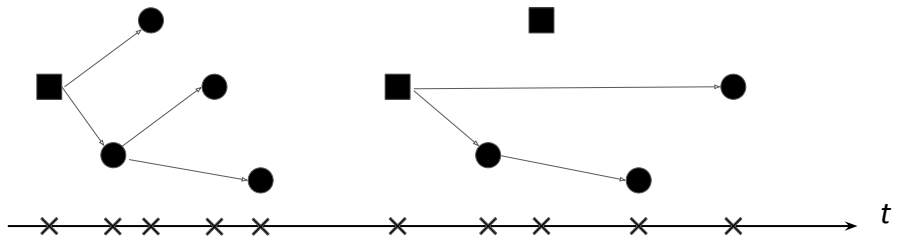


Figure 2.3: An illustration of the immigrant/offspring representation of a Hawkes process. Squares indicate immigrants, circles are offsprings, and the crosses denote the occurrence times.

tinuous time in which each event triggers a Poisson process of successor events. The set of observed events is thereby modelled as a superposition of Poisson processes.

An illustration of the Poisson cluster process representation of a point process, where the clusters are generated by a certain branching structure is shown in Figure 2.3. Here, we distinguish between two types of points, immigrants and offsprings, and have the following definition:

Definition 2.12 *(Rasmussen, 2013, Definition 2.2)*

1. *The immigrants $I = \{t_i\}$ follow a Poisson process.*
2. *Each immigrant $t_i \in I$ generates a cluster C_i, and these clusters are independent.*
3. *A cluster C_i consists of points of generations of order $z = 0, 1, \ldots$ with the following structure: Generation 0 consists simply of the immigrant. Recursively, given the $0, \ldots, z$ generations in C_i, each t_j of generation z generates a Poisson process O_j of offsprings of generation $z + 1$.*
4. *If $t_j \in O_i$, we say that t_j is the offspring of t_i or that t_i is the parent of t_j. We also denote the index of the parent t_i of t_j by $i = \mathsf{pa}(j)$. The branching structure is conveniently represented as $\{u_j\}$, where $u_j = i$ if $t_j \in O_i$ or $u_j = 0$ if t_j is an immigrant.*

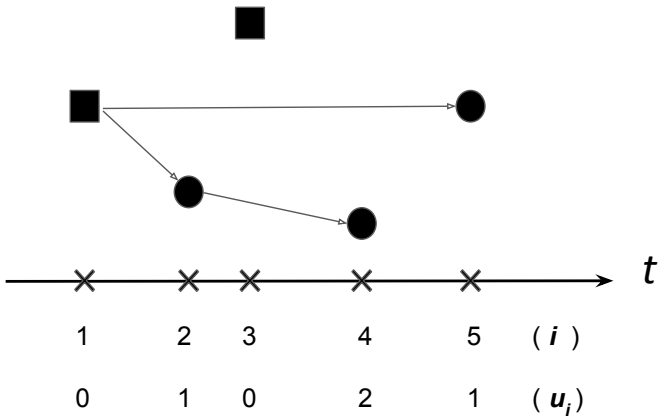


Figure 2.4: An illustration of the branching structure $\{u_i\}$ from Definition 2.12 for a sequence of events. Squares indicate immigrants, circles are offsprings, and the crosses denote the occurrence times.

Let us take a closer look at the branching structure $\{u_i\}$ from Definition 2.12. Immigrants are labelled 0 and each offspring carries the index of its parent. The branching structure is required to completely determine the Poisson cluster process, that is, to decompose it into its independent component clusters and recover the branching tree structure within each cluster. An example illustration of the branching structure for a sequence of events is shown in Figure 2.4.

2.2.4 Likelihood calculations for exponential decay

Consider the case of a Hawkes process where $\lambda^*(t)$ decays exponentially, i.e., $g(u) = \alpha\beta e^{-\beta(t-t_i)}$ and the conditional intensity function is given by

$$\lambda^*(t) = \mu + \sum_{t_i<t} \alpha\beta e^{-\beta(t-t_i)}\,.$$

The parameter $\mu > 0$ is a constant background intensity, $\alpha \in (0,1)$ is the magnitude of excitation that has an upper bound of 1, derived from expression (2.7), to ensure the process remains stationary and $\beta > 0$ is the exponential rate at which the excitation decays over time. Under the immigrant/offspring representation, α can be interpreted as the branching ratio

or the mean number of children of a point and β as the rate of an exponential distribution for the length of the time interval between a child and its parent.

The integrated intensity or the compensator function is

$$\Lambda^*(T) = \int_0^T \lambda^*(u)\,\mathrm{d}u = \mu T - \alpha \sum_{i=1}^{n} \left[\mathrm{e}^{-\beta(T-t_i)} - 1\right] .$$

Let $A(i) = \sum_{j=1}^{i-1} \mathrm{e}^{-\beta(t_i - t_j)}$ with boundary condition $A(1) = 0$, so that

$$A(i) = \mathrm{e}^{-\beta(t_i - t_{i-1})}\,(1 + A(i-1)) .$$

The log-likelihood of the process can be derived from expression (2.1) as

$$\sum_{i=1}^{n} \log(\mu + \alpha\beta A(i)) - \mu T + \alpha \sum_{i=1}^{n} \left[\mathrm{e}^{-\beta(T-t_i)} - 1\right] .$$

2.2.5 Marked Hawkes process

Assume we have observed a marked point process, consisting of event times $\mathbf{t} = \{t_i : t_i \in \mathbb{R}^+ \text{ and } t_i > t_{i-1}\}$ and discrete marks $\mathbf{m} = \{m_i : m_i \in 1, \ldots, M\}\ \forall\ i = 1, \ldots, n$. $M \in \mathbb{N}$ is the number of discrete marks. The marked HP model is most intuitively specified using its mark dependent conditional intensity function $\lambda^*(t, m)$ which for an exponentially decaying intensity is (Rasmussen, 2013, Expression 2.2),

$$\lambda^*(t, m) = \mu\delta_m + \sum_{t_j < t} \alpha\beta \mathrm{e}^{-\beta(t-t_j)} \gamma_{m_j \to m} , \tag{2.9}$$

where the parameter $\mu > 0$ is a constant background intensity and $\delta_m \in [0, 1]$ is the background mark probability for mark m. The parameter $\alpha \in (0, 1)$ is the excitation factor, $\beta > 0$ is the exponential decay rate and $\gamma_{m_j \to m} \in [0, 1]$ is the probability the excitation from an event of mark m_j triggers an event of mark m. Note that by definition $\sum_{m=1}^{M} \delta_m = 1$ and $\sum_{m=1}^{M} \gamma_{m_j \to m} = 1\ \forall\ m_j = 1, \ldots, M$.

Marked Hawkes processes offer a powerful framework to model events whose occurrence rate depends on past occurrences, as they provide insight into the mechanisms governing the process in addition to being able to forecast events. Marked HPs have proven useful in a wide range of applications, for example, in the modelling of earthquakes in Ogata (1998), gang violence in Mohler et al. (2011) or financial market events in Bowsher (2007). However, in applications like the one we are interested in, where events tend not to cluster in time, the marked Hawkes process model is not suitable to be applied as it is.

Chapter 3

Proposed model

Marked Hawkes processes defined in Section 2.2.5 has certain limitations for the modelling of observed event sequences due to their structure. In this chapter, we discuss why these limitations render marked HPs inappropriate for the application we are interested in and then develop a general framework to decouple the joint modelling of the marks and occurrence times. We carry out the derivation of the probability mass function of the marks for the marked Hawkes process model and discuss how its parameters can be interpreted. For the parameter estimation of our proposed model, we develop an EM algorithm by exploiting the branching structure of the process and then discuss the various extensions of our model including a framework for modelling marked spatio-temporal point processes.

3.1 Limitations of the marked Hawkes process model

The characteristic property of the Hawkes process is its self-exciting intensity which naturally leads to clustering of events in time. To illustrate this property, Figure 3.1 shows simulated occurrence times from an unmarked Hawkes Process as well as a homogeneous Poisson process. We also include an instance of observed event times from our football dataset, to show how these two processes compare against the data we are interested to model.

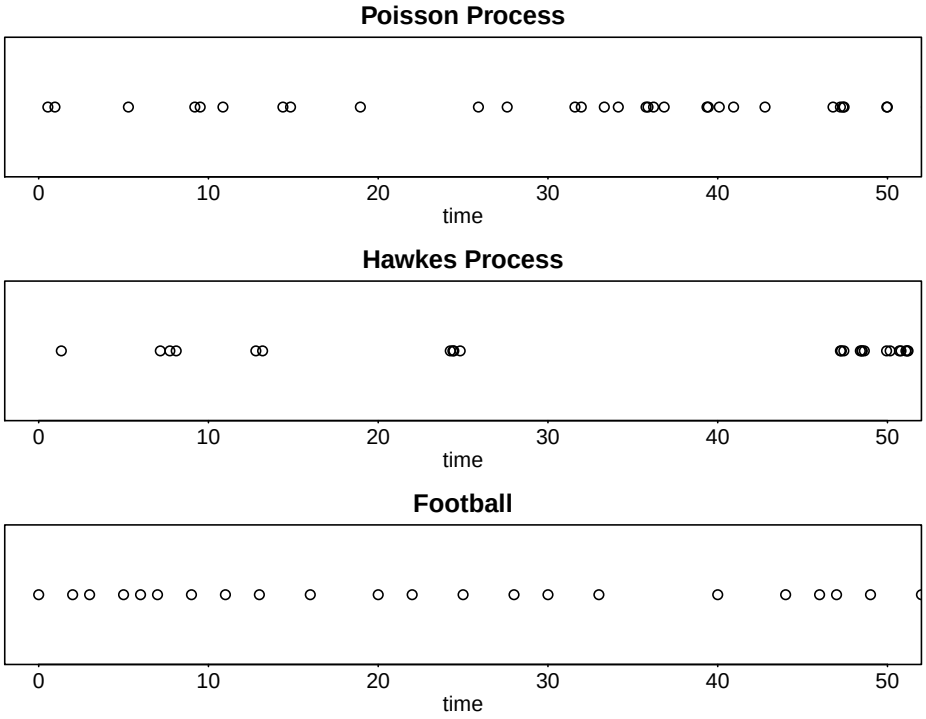


Figure 3.1: Simulated event times from a Poisson process (top) and a Hawkes process (middle) compared against an instance of observed event times from the football dataset (bottom).

The dataset consists of event sequences from football where all match events are recorded (see Section 5.1 for a detailed data description) and we use the event times from a randomly chosen game for our illustration. The parameters for the Hawkes process and the rate for the Poisson process were estimated by fitting the models to the event data in football.

A homogeneous Poisson process has a constant intensity, with exponential inter-arrival times and therefore is a memory-less process that does not exhibit clustering. The Hawkes Process on the other hand is driven by an intensity that increases with every arrival for a short period of time which leads to clustering as seen in Figure 3.1. Crucially, the event times in football appear to be dispersed, even when compared to the Poisson process. This is further evident from the empirical cumulative distribution function

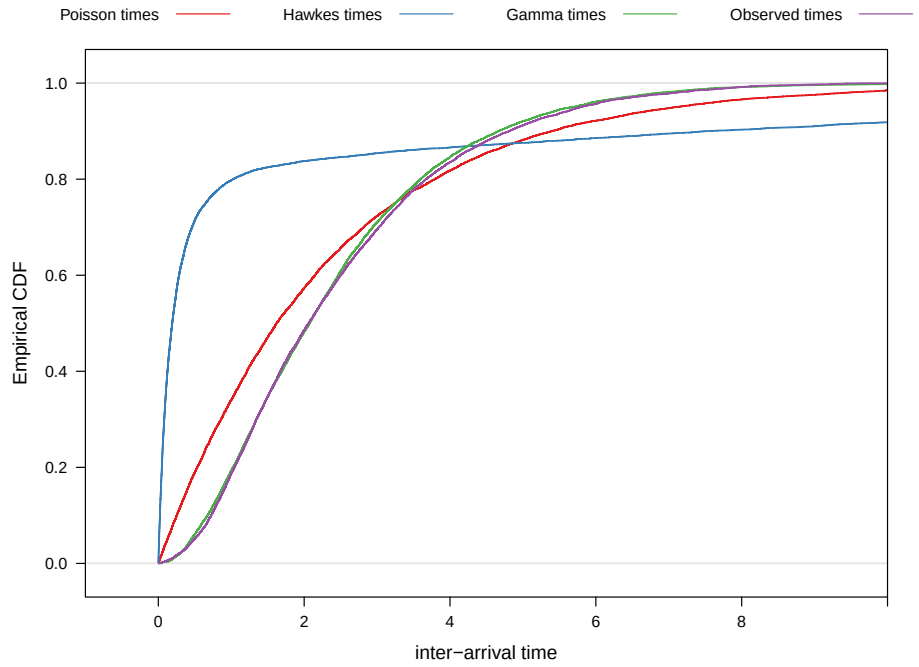


Figure 3.2: Comparing the empirical CDFs of the inter-arrival times of events simulated from a Poisson process (red), a Hawkes process (blue), a Gamma process (green) and observed events in football (purple). ECDFs were computed using 10,000 inter-arrival times in each case.

(ECDF) of the inter-arrival times in Figure 3.2. We cannot analytically compute the CDF of the inter-arrival times for a Hawkes process, and therefore use the empirical CDFs computed using 10,000 simulated inter-arrival times for comparison. The clustering of events in a Hawkes process leads to a higher number of very small as well as very large inter-arrival times, compared to a Poisson process.

Inspection of the empirical CDF of the observed data, led us towards the idea of using a Gamma process to model the inter-arrival times. Figure 3.2 also shows the ECDF from 10,000 simulated inter-arrival times from a Gamma process that was fitted to the observed times in football (see Section 5.7.1 for details). Indeed, we observe that the Gamma process model provides a much better fit to the observed inter-arrival times in football.

We can formally test if the distributions of inter-arrival times are significantly different using the Kolmogorov-Smirnov test (Massey Jr, 1951). The two-sample Kolmogorov–Smirnov test statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| ,$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical cumulative distribution functions of the first and the second sample respectively, sup is the supremum function and n and m are the sample sizes. Intuitively, $D_{n,m}$ quantifies how far apart the empirical CDFs are from each other.

test	$D_{n,m}$	p-value
Hawkes vs Poisson	0.54	$< 10^{-10}$
Football vs Hawkes	0.68	$< 10^{-10}$
Football vs Poisson	0.16	$< 10^{-10}$
Football vs Gamma	0.02	0.16

Table 3.1: Kolmogorov-Smirnov (K-S) test results from the pair-wise comparisons of the distributions of inter-arrival times simulated from a Poisson process, a Hawkes process and a Gamma process with the observed inter-arrival times in football.

The $D_{n,m}$ and p-value from the four pairwise tests are given in Table 3.1, confirming that the Hawkes and Poisson processes are indeed significantly different from the observed times in Football. We used $n = m = 10,000$ samples of inter-arrival times from all the processes to perform the tests. This is evidence that Hawkes processes are not suitable to model events that tend not to cluster in time, like the event-sequences observed in football.

3.2 Decoupling the modelling of times and marks

Traditional models for marked point processes are typically specified using a joint conditional intensity function for the occurrence times and the marks, like marked Hawkes processes in expression (2.9). We find the ex-

citation framework of the marked Hawkes process model appropriate for applications like the event sequences in football, as any event in the sequence is likely to be triggered by one or more of the previous events. The marked Hawkes process model captures the magnitudes all cross-excitations between the various event types as well as the rates at which these excitations decay over time. However, as we have seen, excitation also leads to clustering of events in time and hence, the marked Hawkes process model is not suitable to be applied as it is.

We wish to restrict the characteristic excitation property of marked Hawkes processes exclusively to the modelling of the marks, providing the freedom to specify a different model for the occurrence times. To achieve this, we use the decomposition of a multivariate density function in Cox (1975, Expression 2). Specifically, for a marked point process as defined in Definition 2.6, its full likelihood can always be factorised as

$$\prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) \,, \tag{3.1}$$

where g and f are the conditional probability distribution functions for the times and the marks respectively, and $\boldsymbol{\zeta}, \boldsymbol{\theta}$ are the unknown parameter vectors. Therefore, an alternate approach to specify a marked point process model is to specify the functions g and f separately. The idea is to derive the specification for the marks f from the joint conditional intensity function of a marked Hawkes process model, and then to specify a probability density function for the times g best suited to our application. Thereby, we construct a marked point process model that retains the characteristic properties, like excitation in Hawkes processes, in the model for the marks while avoiding the clustering of event times.

From the definition of the conditional intensity function for a marked point process in expressions (2.2) and (2.3), we have

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\lambda^*(t_i, m_i)}{\sum_{m=1}^{M} \lambda^*(t_i, m)} \,. \tag{3.2}$$

As we restrict ourselves to the case of discrete marks henceforth, we replace the integral in expression (2.3) with a summation over all possible marks in expression (3.2).

To complete the calculations in the case of the marked Hawkes process model, we substitute into expression (3.2) the joint intensity specification from expression (2.9) to get,

$$\begin{aligned} f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) &= \frac{\mu\delta_{m_i} + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}\gamma_{m_j\to m_i}}{\sum_{m=1}^{M}\left[\mu\delta_m + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}\gamma_{m_j\to m}\right]} \\ &= \frac{\mu\delta_{m_i} + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}\gamma_{m_j\to m_i}}{\mu\left[\sum_{m=1}^{M}\delta_m\right] + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}\left[\sum_{m=1}^{M}\gamma_{m_j\to m}\right]} \\ &= \frac{\mu\delta_{m_i} + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}\gamma_{m_j\to m_i}}{\mu + \sum_{t_j<t_i} \alpha\beta \mathrm{e}^{-\beta(t_i-t_j)}}. \end{aligned}$$

Dividing the numerator and denominator by μ and setting $\frac{\alpha\beta}{\mu} = \alpha^*$ we get,

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j<t_i} \alpha^* \mathrm{e}^{-\beta(t_i-t_j)}\gamma_{m_j\to m_i}}{1 + \sum_{t_j<t_i} \alpha^* \mathrm{e}^{-\beta(t_i-t_j)}}. \tag{3.3}$$

We note that the probability mass function f derived in expression (3.3) has rendered some parameters of the original marked Hawkes process model unidentifiable. The parameters μ and α of the original model specified in expression (2.9) described the evolution of the Hawkes process in the time dimension and the sequence of marks specified by f is not sufficient to identify them. Finally, once we specify a probability density function for the event times, $g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta})$, our flexible modelling framework for marked point processes is then complete.

Remark 3.1 *The factorisation in expression (3.1) does not assume the marked point process is separable as defined in Section 2.1.6. The conditional distribution of the mark is allowed to depend on the time of occurrence as well as the history, as seen in expression (3.3). However, we are still able to perform parameter estimation for the functions f and g separately, if they do not share any parameters.*

3.3 Interpreting the model and its parameters

The probability mass function for the marks derived in expression (3.3) has the following interpretation. The mark probability of each event in the sequence is determined by a combined additive effect from a background component and all previous occurrences. The first term in the numerator is the mark probability associated with the background component, while each term in the summation is the contribution from the excitation caused by a previous occurrence in the sequence. The denominator is a normalisation term that ensures the probability mass function sums to 1 over all possible marks.

Background mark probability

The background mark probability $\delta_m \in [0, 1]$ is the probability an event has a mark m if the event is triggered solely by the background component. By definition, we have $\sum_{m=1}^{M} \delta_m = 1$.

Excitation factor

The excitation factor $\alpha^* \geq 0$ is a scaling factor applied to the contributions from the previous occurrences to the event mark probability. Relatively large values of α^* would indicate a stronger dependence of the process on its history, as the contributions from previous occurrences are weighted higher in comparison to the background component.

Decay rate

The decay rate $\beta > 0$ is the exponential rate at which the excitations from previous occurrences decay over time.

Conversion rates

The parameter $\gamma_{m_j \to m_i} \in [0, 1]$ is the probability the excitation from an event of mark m_j triggers an event of mark m_i. In other words, $\gamma_{m_j \to m_i}$ can be viewed as the conversion rate for the transition $m_j \to m_i$. By definition, we have $\sum_{m=1}^{M} \gamma_{m_j \to m} = 1 \; \forall \; m_j = 1, \ldots, M$.

In summary, the specification for the marks in expression (3.3) captures all cross-excitations between the various marks as well as the rate at which these excitations decay over time, which are the crucial features of the marked Hawkes process model we wished to retain.

3.4 Parameter estimation via Expectation-Maximisation

Assuming the functions f and g do not share any parameters, we can perform parameter estimation for the sequence of the marks and the occurrence times separately. In the framework introduced in Section 3.2, the model for the event times g is left open for choice and typically, we use a simple model like a Gamma process for which the parameter estimation is trivial. Therefore, we focus on the parameter estimation of the model for the marks specified by f in expression (3.3).

To summarise, we are left with a multi-class classification problem of modelling the random mark sequence $\boldsymbol{m} = \{m_i\}_{i=1}^{n}$ as specified by (3.3) given the time sequence $\boldsymbol{t} = \{t_i\}_{i=1}^{n}$ and need to estimate the model parameters $\boldsymbol{\theta}$.

The likelihood for a complete mark sequence from (3.3) is

$$\prod_{i=1}^{n} f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}}, \tag{3.4}$$

where the transformation $\alpha = \log(\alpha^*)$ is applied to ensure that $\alpha^* \geq 0$.

The direct maximisation of the likelihood in expression (3.4) suffers from the same issues faced by the marked Hawkes process model. As discussed

by Veen and Schoenberg (2008), the difficulties arise due to the term in the likelihood involving sums over previous points and the fact that the log-likelihood can be nearly flat in large regions of the parameter space. Veen and Schoenberg (2008) proposed an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) as an efficient alternative for the parameter estimation of Hawkes processes, which we adapt to our model as follows.

We introduce a latent quantity u_i, which indicates whether the i-th event came from the background ($u_i = 0$) or was triggered by a previous event with index j ($u_i = j$). The latent quantity u_i is the branching structure from the immigrant/offspring representation discussed in Section 2.2.3. If the branching structure u_i is assumed to be known, the complete-data log-likelihood for a parameter vector $\boldsymbol{\theta}$ is

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) = &\sum_{i=1}^{n} \mathbb{1}(u_i = 0) \log\left(\delta_{m_i}\right) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbb{1}(u_i = j) \log\left(\mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}\right) \\ &- \sum_{i=1}^{n} \log\left(1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}\right),\end{aligned}$$

where $\mathbb{1}(.)$ is the indicator function, which takes the value 1 when its argument holds and zero otherwise. The branching structure simplifies the log-likelihood, as the mark probability for each event is determined only from its trigger (the background or a previous event). This is similar to the common EM approach to mixture models, where the latent variables indicate the underlying distribution from which each data point arose.

To complete the E step, we take the expectation of $\ell_c(\boldsymbol{\theta})$. This requires estimating the branching structure probabilities $\mathbb{P}(u_i = j \mid \mathcal{F}_{t_i}) = \mathbb{E}[\mathbb{1}(u_i = j) \mid \mathcal{F}_{t_i}]$ for all i, j, based on the parameter values $\hat{\boldsymbol{\theta}}$ of the current iteration.

We can calculate these probabilities as

$$\mathbb{P}(u_i = 0 \mid \mathcal{F}_{t_i}) = \frac{\delta_{m_i}}{\delta_{m_i} + \sum_{t_k < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_k)} \gamma_{m_k \to m_i}},$$
$$\mathbb{P}(u_i = j \mid \mathcal{F}_{t_i}) = \begin{cases} \frac{\mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}}{\delta_{m_i} + \sum_{t_k < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_k)} \gamma_{m_k \to m_i}} & \text{for } t_j < t_i \\ 0 & \text{for } t_j \geq t_i \end{cases}. \tag{3.5}$$

This leads to the expected complete-data log-likelihood which is then maximised in the M step

$$\begin{aligned} \mathbb{E}[\ell_c(\boldsymbol{\theta})] = &\sum_{i=1}^{n} \mathbb{P}(u_i = 0 \mid \mathcal{F}_{t_i}) \log(\delta_{m_i}) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbb{P}(u_i = j \mid \mathcal{F}_{t_i}) \log\left(\mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}\right) \\ &- \sum_{i=1}^{n} \log\left(1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}\right). \end{aligned}$$

The current parameter estimates $\hat{\boldsymbol{\theta}}$ are updated at the end of the M step and the procedure returns to the E step, estimating new triggering probabilities, and repeats until the log-likelihood converges.

3.5 Model extensions

The probability mass function for the marks as specified in expression (3.4) can be extended in many ways that could potentially result in a better fit to the data. In this section we discuss approaches that we believe are most likely to lead to significant improvements in real-world applications, as shown in Section 5.8.

3.5.1 Covariate dependent conversion rates

Often, there may be additional covariates associated with the events in the data available to us. Event specific covariates can be incorporated in the conversion rate parameters using a baseline-category logit specification (see,

for example, Agresti, 2007, Section 6.1). In the application for this thesis, football team information can be incorporated as

$$\log\left(\frac{\gamma_{m_j\to m}(h)}{\gamma_{m_j\to M}(h)}\right) = \varphi_{m_j\to m} + \omega_{h,m} \qquad \forall\, m \in 1,\ldots,M-1\,, \tag{3.6}$$

where φ is the baseline conversion parameter and h is the team in possession of the ball attempting the event conversion. The parameter ω is then interpreted as the relative ability of a team to complete a conversion to an event of mark m.

3.5.2 Event dependent decay rates

The decay rate β in expression (3.4) is fixed irrespective of the mark that triggered the excitation or the mark that is being excited. In real-world scenarios however, the excitation effects may vary across different mark pairs, some effects persisting over a long duration while others short. To capture such scenarios we let the excitation decay rate β depend on the pair of marks involved in the excitation. The resulting probability mass function for marks is

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j<t_i} \mathrm{e}^{\alpha-\beta_{m_j\to m_i}(t_i-t_j)}\gamma_{m_j\to m_i}}{\sum_{m=1}^{M}\left[\delta_m + \sum_{t_j<t_i} \mathrm{e}^{\alpha-\beta_{m_j\to m}(t_i-t_j)}\gamma_{m_j\to m}\right]}\,, \tag{3.7}$$

where $\beta_{m_j\to m_i}$ is the exponential decay rate of the excitation caused by an event of mark m_j on an event of mark m_i. Expression (3.7) allows dependence between event types over arbitrary lengths of time and can provide valuable insight into the dynamics of the underlying process. However the matrix parameterisation introduces M^2 excitation decay rates that could potentially make estimation of the parameters challenging.

An alternate approach to allow the excitation decay rate β to depend on the pair of marks involved in the excitation is to use a product of two vectors

parameterisation. The resulting probability mass function for marks is

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta^1_{m_j} \beta^2_{m_i} (t_i - t_j)} \gamma_{m_j \to m_i}}{\sum_{m=1}^{M} \left[\delta_m + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta^1_{m_j} \beta^2_{m_i} (t_i - t_j)} \gamma_{m_j \to m} \right]}, \tag{3.8}$$

where $\beta^1_{m_j}$ and $\beta^2_{m_i}$ are the contributions to the decay rate of the excitation caused by an event of mark m_j on an event of mark m_i. The parameterisation in expression (3.8) introduces only $2M$ excitation decay rates and offers a more computationally feasible option. However this comes at the expense of flexibility in capturing the dependence between the different event types and for this reason we preferred to implement the parameterisation in expression (3.7) for the application presented in Chapter 5.

3.6 Marked spatio-temporal point processes

Until now we described a modelling framework for marked point processes that consisted of event occurrence times and discrete marks. In addition to the times and the marks, if there is a location associated with each event, then the process can be modelled as a marked spatio-temporal point process. A marked spatio-temporal point process $\boldsymbol{X}$, consists of event times $\boldsymbol{t} = \{t_i : t_i \in \mathbb{R} \text{ and } t_i > t_{i-1}\}$, locations $\boldsymbol{z} = \{z_i : z_i \in \mathcal{Z} \subseteq \mathbb{R}^d\}$ and marks $\boldsymbol{m} = \{m_i : m_i \in 1, \ldots, M\}\ \forall\ i = 1, \ldots, n$. $M \in \mathbb{N}$ is the number of discrete marks.

Definition 3.2 *We define the history or filtration $\mathcal{F}_t$ at time t of the process $\boldsymbol{X}$ as $\mathcal{F}_t = \{(t_j, z_j, m_j) : t_j \in \boldsymbol{t}, z_j \in \boldsymbol{z}, m_j \in \boldsymbol{m} \text{ and } t_j \leq t\}$.*

Similar to marked point processes, the full likelihood of a marked spatio-temporal point process can be factorised as

$$\prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} h(z_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\eta}) \prod_{i=1}^{n} f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}), \tag{3.9}$$

where g, h and f are the probability distribution functions for the times, the locations and the marks respectively, and $\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\theta}$ are the corresponding unknown parameter vectors.

A natural way to incorporate location information in the probability mass function for the marks f, is to allow the parameters to vary according to the event location. The probability mass function for marks in expression (3.4) can rewritten as

$$f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i}(z_i) + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(z_i)(t_i - t_j)} \gamma_{m_j \to m_i}(z_i)}{\sum_{m=1}^{M} \left[\delta_m(z_i) + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(z_i)(t_i - t_j)} \gamma_{m_j \to m}(z_i) \right]}, \tag{3.10}$$

where $\delta_{m_i}(z_i)$ is the location dependent background mark probability of mark m_i. Similarly, $\beta(z_i)$ and $\gamma_{m_j \to m}(z_i)$ are the location dependent decay and conversion rates respectively.

Finally, in addition to g, we also need to specify h, a model for the locations, best suited to our application to complete the modelling framework for marked spatio-temporal point processes. Parameter estimation via the EM algorithm detailed in Section 3.4, can be easily adapted to the extensions of the modelling framework specified in expressions (3.7) and (3.10) by using the appropriate complete data log-likelihood and branching structure probabilities.

Chapter 4

Bayesian inference

One of our primary research goals is to be able to do online inference, that would enable us to make predictions in real-time. The Bayesian paradigm of updating one's beliefs based on new information is well suited to such a task (see, for example, Bernardo and Smith, 2007). In general, given an observed data sample $\mathbf{X}$ and a data model with parameter vector $\boldsymbol{\theta}$, Bayesian inference involves the application of Bayes' rule to relate the posterior probability distribution $p(\boldsymbol{\theta} \mid \mathbf{X})$ of the parameter vector $\boldsymbol{\theta}$ with the likelihood $p(\mathbf{X} \mid \boldsymbol{\theta})$ and prior distribution $p(\boldsymbol{\theta})$ in the following way

$$p(\boldsymbol{\theta} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \,.$$

The posterior predictive distribution of a new data point $\tilde{\mathbf{X}}$ is then

$$p(\tilde{\mathbf{X}} \mid \mathbf{X}) = \int p(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{X}) \, \mathrm{d}\boldsymbol{\theta} \,.$$

If computing this integral analytically is infeasible, we use a Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X})$. Then, the posterior predictive distribution can be approximated by

$$p(\tilde{\mathbf{X}} \mid \mathbf{X}) \approx \frac{1}{R} \sum_{k=1}^{R} p(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}_k) \,,$$

where $\boldsymbol{\theta}_k$ is one of R parameter samples from its posterior distribution. By returning a predictive distribution in this way, Bayesian inference quantifies the uncertainty associated with the prediction.

In this chapter, we present a Bayesian framework for the inference and prediction of marked point processes. We first specify the full likelihood for a collection of event sequences and the prior distributions for the model parameters. We then discuss the Hamiltonian Monte Carlo algorithm for sampling from the posterior distribution including its software implementation. Finally, we provide details on the methods for model evaluation including the set-up of a simulation framework and the performance measures used for its validation.

4.1 Model specification

We develop a Bayesian modelling framework for a collection of event sequences, where each sequence is modelled using the decoupled model for marked point processes proposed in Chapter 3.

4.1.1 Likelihood

From expression (3.1), for a total of S sequences, the likelihood is

$$\prod_{s=1}^{S} \left[\prod_{i=1}^{n_s} g(t_{s,i} \mid \mathcal{F}_{t_{s,i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n_s} f(m_{s,i} \mid t_{s,i}, \mathcal{F}_{t_{s,i-1}}; \boldsymbol{\theta}) \right], \tag{4.1}$$

where n_s is the number of events in the sequence s, $t_{s,i}$ and $m_{s,i}$ are the occurrence time and mark of the i-th event in the sequence s respectively. Note that, all sequences in the collection are modelled independently of each other and therefore we suppress the subscript s henceforth for notational convenience.

The probability density function for the occurrence times is set to

$$g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) = p(t_i - t_{i-1} \mid m_{i-1}, \boldsymbol{a}, \boldsymbol{b})$$
$$t_i - t_{i-1} \mid m_{i-1}, \boldsymbol{a}, \boldsymbol{b} \sim \textbf{Gamma}[a(m_{i-1}), b(m_{i-1})]\,, \tag{4.2}$$

where the time to next event in each sequence is modelled using a gamma distribution with shape and rate parameters that depend on the mark of the last observed event. This specific choice for g is purely based on its suitability for our application and is justified in Section 5.5. We assume the underlying process that generates the event sequence is observed from the time of occurrence of the first event and set $t_1 = 0$ and $g(t_1) = 1$.

We use the probability mass function, derived in expression (3.4), as the specification for the marks

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}}\,, \tag{4.3}$$

including the baseline-category logit specification, from expression (3.6), to incorporate event-specific covariates in the conversion rate parameters as

$$\log\left(\frac{\gamma_{m_j \to m}(h)}{\gamma_{m_j \to M}(h)}\right) = \varphi_{m_j \to m} + \omega_{h,m} \qquad \forall\, m \in 1, \ldots, M-1\,. \tag{4.4}$$

4.1.2 Graphical model

Figure 4.1 provides a graphical representation of the model. As shown, the collection of observed event sequences are modelled as independent processes with shared parameters $\alpha, \beta, \boldsymbol{\delta}, \boldsymbol{\zeta}, \boldsymbol{\varphi}$. Process specific information is accounted for in the parameter vector $\boldsymbol{\omega}$. Furthermore, the process-level graph illustrates the conditioning of variables within each sequence, highlighting how for each event, the occurrence time is first modelled given the history up-to the current event. The event mark is then modelled given its occurrence time and the history.

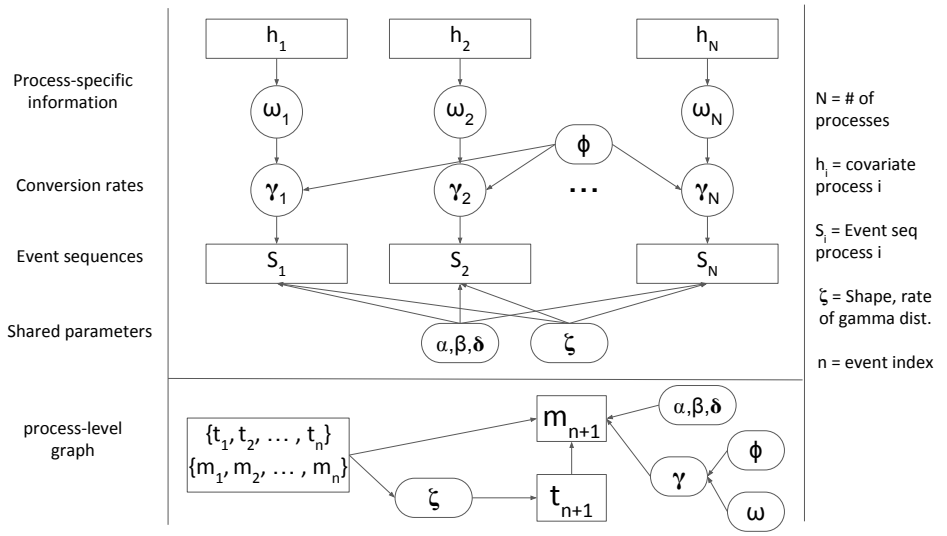


Figure 4.1: Graphical model showing conditional dependencies in the Bayesian model specification in Section 4.1.

4.1.3 Prior distributions

The shape and rate parameters of the Gamma distribution for inter-arrival times are non-negative and assigned exponential priors with shared rate hyper-parameters a' and b',

$$\boldsymbol{a} \sim \mathbf{Exp}(a')\,, \boldsymbol{b} \sim \mathbf{Exp}(b')\,. \tag{4.5}$$

The background mark probability vector δ has a multinomial distribution, i.e., $\sum_{m=1}^{M} \delta_m = 1$. We assign a Dirichlet prior on δ which is conjugate to the multinomial distribution with a concentration hyper-parameter δ', i.e.

$$\delta \sim \mathbf{Dirichlet}(\delta')\,.$$

The excitation factor α is unbounded and assigned a normal prior with a hyper-parameter σ_α,

$$\alpha \sim \mathbf{N}(0, \sigma_\alpha)\,.$$

The decay rate parameter $\beta > 0$ being non-negative, is assigned an exponential prior with a rate hyper-parameter β' ,

$$\beta \sim \mathbf{Exp}(\beta') \,.$$

We assign a Normal shrinkage prior with non-centred parameterisation and a shared hyper-parameter σ_γ on the unbounded parameters of the baseline-category logit model parameters,

$$\boldsymbol{\varphi}, \boldsymbol{\omega} \sim \mathbf{N}(0, \sigma_\gamma) \,.$$

Finally, we assign non-informative hyper-priors for all hyper-parameters in our model,

$$p(a', b', \boldsymbol{\delta'}, \sigma_\alpha, \beta', \sigma_\gamma) \propto 1$$

4.1.4 Impact of prior distributions

The prior distributions for the model parameters specified in Section 4.1.3 fall into the category of non-informative or weakly informative priors. The choices are out of convenience to allow the posterior distributions to concentrate around the maximum likelihood estimate. We wish to have prior distributions that are flat and spanning a wider region of the parameter space as compared to their corresponding posterior distributions. A convenient way to formally test sensitivity with respect to the prior specification is to inspect the ratio of the variance of the prior distribution to the variance of the posterior distribution for each parameter (Millar, 2004). Values much larger than 1 would indicate that the priors are flat compared to the posterior.

The model introduces $\mathcal{O}(M^2)$ number of parameters, making learning the parameters of this model a potentially challenging task as well as increasing the chance of over-fitting. In Section 5.6, we propose an algorithm based on

association rules to reduce the number of estimated parameters. However, such model complexity could also be dealt with using an appropriate prior specification like the spike and slab variable selection strategy (Ishwaran et al., 2005). We believe the handling of model complexity using priors to be a valuable extension to the proposed framework that could be explored as part of the future work.

4.2 Posterior sampling algorithm

Posterior sampling for the Bayesian model as specified in Section 4.1 can be performed using Gibbs sampling (Geman and Geman, 1984). However, in cases like ours, where the posterior distribution is high-dimensional, typical inference methods like the Metropolis algorithm (Metropolis et al., 1953) and Gibbs sampling suffer due to their inherent random walk behaviour (Neal, 1993). We perform posterior sampling via a variant of the Hamiltonian Monte Carlo algorithm, originally proposed by Duane et al. (1987), to obtain samples from the posterior distribution of the model parameters. Borrowing a concept from physics, the Hamiltonian Monte Carlo (HMC) algorithm is able to suppress the random walk behaviour and explore the posterior distribution efficiently. In this section, we first briefly present how the Gibbs sampling method can be applied to our model even though it proved to be computationally infeasible ultimately. We then discuss the key elements of the HMC algorithm and its adaptive variant the no-U-turn sampler (NUTS), along with details of their implementation in the software package Stan by Stan Development Team (2020).

The goal of any posterior sampling algorithm is to draw samples from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X})$ of the parameter vector $\boldsymbol{\theta}$ given the data $\mathbf{X}$.

4.2.1 Gibbs sampling

The main idea in Gibbs sampling (Geman and Geman, 1984) is to generate posterior samples by sweeping through each variable to sample from its full conditional distribution with the remaining variables fixed to their current values. This process repeats until the sample values converge to the true posterior joint distribution.

Assuming the functions f and g do not share any parameters in the likelihood in expression (4.1), we can perform parameter estimation for the sequence of the marks and the occurrence times separately. The model for the event times g is a Gamma process for which the parameter estimation is trivial. Therefore, we focus on the parameter estimation of the model for the marks specified by f in expression (4.3).

We are left with a multi-class classification problem of modelling the random mark sequence $\boldsymbol{m} = \{m_i\}_{i=1}^n$ as specified by (4.3) given the time sequence $\boldsymbol{t} = \{t_i\}_{i=1}^n$ and need to estimate the model parameters $\boldsymbol{\theta}$.

Data augmentation

We begin by augmenting the data with the unobserved branching structure $\boldsymbol{u}$ discussed in Section 2.2.3, which indicates whether the i-th event came from the background ($u_i = 0$) or was triggered by a previous event with index j ($u_i = j$). The joint posterior density is

$$p(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{m}) \propto p(\boldsymbol{m}, \boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \,.$$

For the model parameters in $\boldsymbol{\theta}$, we use the prior specification from Section 4.1.3. The branching structure simplifies the probability mass function for the marks f in expression (4.3), as the mark probability for each event is determined only from its trigger (the background or a previous event). The

joint likelihood of the marks and the latent branching structure is

$$p(m_i, u_i = 0 \mid \boldsymbol{\theta}) = \frac{\delta_{m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}}$$
$$p(m_i, u_i = j \mid \boldsymbol{\theta}) = \begin{cases} \frac{\gamma_{m_j \to m_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}}{1 + \sum_{t_k < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_k)}} & \text{for } t_j < t_i \\ 0 & \text{for } t_j \geq t_i \end{cases} . \tag{4.6}$$

Full conditionals

The full conditional distribution of the model parameters can be calculated from the full joint distribution as

$$p(\boldsymbol{\theta} \mid \boldsymbol{m}, \boldsymbol{u}) \propto p(\boldsymbol{m}, \boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) ,$$

where the joint likelihood $p(\boldsymbol{m}, \boldsymbol{u} \mid \boldsymbol{\theta})$ is given in expression (4.6). Specifically, we use a Metropolis within Gibbs sampler to sample $\boldsymbol{\theta}$ from its full-conditional density $p(\boldsymbol{\theta} \mid \boldsymbol{m}, \boldsymbol{u})$.

The full conditional distribution of the branching structure is

$$p(\boldsymbol{u} \mid \boldsymbol{m}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{m}, \boldsymbol{u} \mid \boldsymbol{\theta})}{p(\boldsymbol{m} \mid \boldsymbol{\theta})} ,$$

where $p(\boldsymbol{m} \mid \boldsymbol{\theta})$ is the probability density function of the marks from expression (4.3).

Posterior samples are then simulated by sweeping through all the full conditionals, one variable at a time.

4.2.2 Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo algorithm adds an auxiliary momentum variable ρ_j for each component of the parameter vector θ_j (Neal, 2011). The posterior density $p(\boldsymbol{\theta} \mid \mathbf{X})$ is augmented with an independent distribution $p(\boldsymbol{\rho})$ of the momentum variables $\boldsymbol{\rho}$ to define a join density

$$p(\boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{X}) \propto p(\boldsymbol{\rho}) p(\boldsymbol{\theta} \mid \mathbf{X}) ,$$

from which both $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ are sampled together.

As a first step in the HMC algorithm, the momentum variables $\boldsymbol{\rho}$ are sampled from a multivariate Normal distribution that does not depend on $\boldsymbol{\theta}$

$$\boldsymbol{\rho} \sim \mathsf{MultiNormal}(0, M)\,,$$

where the covariance matrix M, also called the *mass* matrix, is typically chosen to be diagonal, meaning the components ρ_j are independent.

Hamiltonian dynamics

The augmented density $p(\boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{X})$ can be interpreted in physical terms as a Hamiltonian system where $\boldsymbol{\theta} \in \mathbb{R}^{\mathsf{d}}$ denotes the position of a particle in d-dimensional space and $\boldsymbol{\rho}$ its momentum. The Hamiltonian is defined as (Betancourt, 2017),

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{\rho}) &= -\log p(\boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{X}) \\ &= -\log p(\boldsymbol{\rho}) - \log p(\boldsymbol{\theta} \mid \mathbf{X}) \\ &= K(\boldsymbol{\rho}) + V(\boldsymbol{\theta})\,, \end{aligned}$$

where $K(\boldsymbol{\rho}) = -\log p(\boldsymbol{\rho})$ is the kinetic energy of the particle and $V(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \mid \mathbf{X})$ is the potential energy function. The joint system $(\boldsymbol{\theta}, \boldsymbol{\rho})$ evolves according to Hamilton's equations,

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= +\frac{\partial K}{\partial \boldsymbol{\rho}} \\ \frac{d\boldsymbol{\rho}}{dt} &= -\frac{\partial V}{\partial \boldsymbol{\theta}}\,. \end{aligned}$$

In the second step of HMC, a leapfrog integrator solves Hamilton's equations by taking L discrete steps of some small time interval ε. Each leapfrog

step consists of alternating half-step updates of $\boldsymbol{\rho}$ and full-step updates of $\boldsymbol{\theta}$

$$\begin{aligned}\boldsymbol{\rho} &\leftarrow \boldsymbol{\rho} - \frac{\varepsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}} \\ \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \varepsilon M^{-1}\boldsymbol{\rho} \\ \boldsymbol{\rho} &\leftarrow \boldsymbol{\rho} - \frac{\varepsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}}\,.\end{aligned}$$

At the end of L leapfrog steps, the resulting state is denoted as $(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$.

Metropolis acceptance step

In the final step of HMC, a Metropolis acceptance step is applied to account for the numerical errors during the leapfrog integration procedure. The probability of accepting the proposal $(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$ starting from the current state $(\boldsymbol{\theta}, \boldsymbol{\rho})$ is

$$\min\left(\frac{p(\boldsymbol{\theta}^*, \boldsymbol{\rho}^* \mid \mathbf{X})}{p(\boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{X})}, 1\right)\,.$$

If the proposal is rejected, the current state $(\boldsymbol{\theta}, \boldsymbol{\rho})$ is returned and used to initialise the next iteration.

4.2.3 HMC implementation in Stan

Stan automatically applies the HMC algorithm given a Bayesian model to generate parameter samples from the posterior distribution. Stan implements the algorithm in the following key steps,

1. Input of data, model and parameter initialisation.
2. Calculation of the log posterior density and its gradients.
3. Calibration of tuning parameters (e.g. number of leapfrog steps) in a warm-up phase.
4. Implementation of the No-U-Turn Sampler (NUTS) to generate samples from the posterior distribution.

The No-U-Turn Sampler (Hoffman and Gelman, 2014) is an adaptive variant of the HMC algorithm in which the tuning parameters are automatically determined.

Data and model input

A Bayesian model specified in a Stan program consist of variable type declarations and statements in blocks corresponding to the purpose of the variable: functions, data, parameter and transformed parameter. Figure 4.2 provides a snippet of the Stan program that implements the Bayesian model specified in Section 4.1. The process_logl function calculates the log-likelihood and is defined in the functions block. The arrays of observed times and marks along with its meta data are defined in the data block. The parameters are declared in the parameters block and the variable transformations are defined in the transformed parameters block. The model block consists of statements defining the choice of prior distributions for the parameters and the call to the log-likelihood function. In addition to the data, parameters, and model statements, at execution, a Stan program also requires the number of chains, the number of iterations and starting values for each parameter per chain.

```
functions { // Function to compute the log-likehood
  vector process_logl(vector params, real[] time_array, int[] marks_array) {

    int L = size(time_array);
    int   i = 1;
    real ll = 0;

    while( i <= L ){ // add up event-wise log-likelihood for marks

        ll += < // code to compute log-likelihood for the i-th mark />;

        i += 1;
    }

    return [ll]';
  }
}
```

```
data{
  int<lower=1> M; // Number of distinct marks
  real<lower=0> times[]; // observed times
  int<lower=0> marks[]; // observed marks
}

parameters{
  simplex[M] delta; // unknown delta
  real alpha; // unknown alpha
  real<lower=0> beta; // unknown decay rates
  matrix[M, M-1] theta; // unknown baseline conversion rates
}

transformed parameters{
  matrix[M, M-1] gamma; // transformed baseline conversion rates

  <// code to transform conversion rates using baseline logit specification //>;
}

model{
  // Prior specification
  delta ~ dirichlet(rep_vector(1.0, M));
  alpha ~ normal(0, 10);
  beta ~ exponential(0.01);
  theta ~ normal(0, 10);

  // Parallel call to log-likelihood function for the sequences of marks
  target += sum(map_rect(process_logl, params, mu_process, times, marks));
}
```

Figure 4.2: Stan program that implements the Bayesian model specified in Section 4.1.

Automatic parameter tuning

The HMC algorithm requires three parameters to be set,

- *mass* matrix M,
- number of leapfrog steps L, and
- discretisation time ε.

By default Stan sets M equal to the diagonal estimate of the inverse posterior covariance matrix $(\mathsf{var}(\boldsymbol{\theta} \mid \mathbf{X}))^{-1}$ computed at the end of a warm-up phase.

If M^{-1} is a poor estimate of the posterior covariance, ε must be kept small to maintain arithmetic precision, lowering the overall efficiency of the sampling algorithm (Gelman et al., 2013).

Stan implements the No-U-Turn Sampler (NUTS) where the number of leapfrog steps L is adaptively determined at each iteration. Intuitively, the trajectory of leapfrog steps in the NUTS algorithm continues until a balance condition is satisfied, before it starts to turn around (Hoffman and Gelman, 2014). NUTS also provides a method for adapting the discretisation time ε dynamically based on primal-dual averaging (Nesterov, 2009).

4.3 Sequential updating via importance sampling

Let $\boldsymbol{y}_k = \{\zeta_k, \boldsymbol{\theta}_k\}$ for $k = 1, \ldots, R$ be a sample of the posterior parameter vector after training the model on training data $\mathbf{X}$. And let $\mathbf{X}' = \{t_i, m_i\}$ for $i = 1, \ldots, n$ be the new test data consisting of the event history of a sequence not in $\mathbf{X}$, till some intermediate time T during which n events occurred.

When presented with the new data $\mathbf{X}'$, we can update the parameter samples $\boldsymbol{y}_k$ using importance sampling (Chopin, 2002). The idea is to resample with replacement from the posterior samples $\boldsymbol{y}_k$ using unequal weights that are proportional to the ratio of the likelihood of the complete data ($\mathbf{X}$ and $\mathbf{X}'$) to the likelihood of the training data ($\mathbf{X}$) only. For each sample $\boldsymbol{y}_k$ we calculate its weight w_k as,

$$\begin{aligned} \widetilde{w_k} &= \frac{p(\mathbf{X}, \mathbf{X}' \mid \boldsymbol{y}_k)}{p(X \mid \boldsymbol{y}_k)} \\ w_k &= \frac{\widetilde{w_k}}{\sum_k \widetilde{w_k}} . \end{aligned} \tag{4.7}$$

We then resample R times with replacement samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_R$ with probabilities $w_1, \ldots, w_R$ to get the updated samples $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_R$. If a single w_k turns out to be vastly larger than all the others, we may end up with many replications of the same sample. A diagnostic than can inform us if the weights

are problematic is the effective sample size (Kong, 1992), which is calculated as

$$n_{\mathrm{e}} = \frac{1}{\sum_{k=1}^{R} w_k^2},$$

and $n_{\mathrm{e}} << R$ indicates that the weights are highly imbalanced.

4.4 Model evaluation

In this section, we discuss two approaches to evaluate the accuracy of the Bayesian model for marked point processes. The first approach detailed in Section 4.4.1 relies on using the log-likelihood of the test data evaluated at the posterior parameter samples to compute a log score which can be used for model comparison. In the second approach discussed in Section 4.4.2, we simulate a sequence of events in a specified interval given its history, and then validate the simulated event counts against the observed counts. Prior to simulation, the posterior samples are updated given the new data using the sequential updating method in Section 4.3.

4.4.1 Log point-wise predictive density

A straightforward method to evaluate the predictive accuracy of a model is to use the log point-wise predictive density computed on the test data. The log point-wise predictive density of the test data $\mathbf{X}'$ can be computed using the samples from the posterior as (Vehtari et al., 2017)

$$\widehat{lpd} = \sum_{i=1}^{n} \log \left(\frac{1}{R} \sum_{k=1}^{R} p\left(t_i, m_i \mid \boldsymbol{y}_k\right) \right), \tag{4.8}$$

where $p\left(t_i, m_i \mid \boldsymbol{y}_k\right)$ is the likelihood of the i-th event in the process evaluated at sample $\boldsymbol{y}_k$.

The $\widehat{lpd}$ values calculated based on different models can be used to compare them, with larger values indicating better predictive accuracy.

4.4.2 Simulation based validation

Given training data $\mathbf{X}$ and a sequence of n events $\mathbf{X}'$ from the test data, the posterior predictive distribution for the $(n+1)$th event, the pair of occurrence time t_{n+1} and mark m_{n+1} is

$$\begin{aligned} p(t_{n+1}, m_{n+1} \mid \mathbf{X}, \mathbf{X}') &= \int p(t_{n+1}, m_{n+1} \mid \boldsymbol{\zeta}, \boldsymbol{\theta}, \mathbf{X}, \mathbf{X}') p(\boldsymbol{\zeta}, \boldsymbol{\theta} \mid \mathbf{X}, \mathbf{X}') \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\zeta} \\ &\approx \frac{1}{R} \sum_{k=1}^{R} p(t_{n+1}, m_{n+1} \mid \mathbf{X}, \mathbf{X}', \boldsymbol{q}_k), \end{aligned}$$

where $\boldsymbol{q}_k$ is a sample from the posterior parameter distribution for $k = 1, \ldots, R$ after updating.

For model evaluation, we simulate a sequence of events in some predefined interval and compare against the observed truth. For each event sequence in the test data, we first compute the updated posterior samples $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_R$ given the new event history till some time T and then generate Q simulations per sample in the interval $(T, T+d)$, where is T is the time at which each simulation is started and d is the duration of the simulation interval.

Each simulation is carried out iteratively as follows; we first simulate the occurrence time of next event given history and then its mark given time and history. This generated pair of (time, mark) is then added to the history as the most recent event. The simulation is stopped when the time exceeds $T+d$. Finally, for each event sequence in the test set, we validate the event counts from the $R \times Q$ simulations against the observed counts in the simulation interval using an appropriate performance measure.

4.4.3 Performance measures

There are a number of scoring rules that can be used as performance measures in the evaluation of probabilistic forecasts for count data (Czado et al., 2009). Scoring rules provide a numerical score based on the observation x and the predictive probability distribution P. Let us denote the predicted

probability mass function by $(p_k)_{k=0}^{\infty}$ and the predicted cumulative distribution function by $(P_k)_{k=0}^{\infty}$. Scoring rules are denoted by $s(P, x)$ that the forecaster tries to minimise and are said to be *proper* if the lowest score is achieved by the true probability distribution.

1. **Logarithmic Score**: The logarithmic score is

$$\mathrm{s}(P, x) = -\log p_x \,.$$

   where p_x is the predicted probability mass at the observation x.

2. **Brier Score**: The Brier score is

$$\mathrm{s}(P, x) = -2p_x + \|p\|^2 \,,$$

   where $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$.

3. **Spherical Score**: The spherical score is

$$\mathrm{s}(P, x) = -\frac{p_x}{\|p\|} \,.$$

4. **Ranked Probability Score**: The ranked probability score (RPS) is

$$\mathrm{s}(P, x) = \sum_{k=0}^{\infty} \{P_k - \mathbb{1}(x \leq k)\} \,.$$

5. **Squared Error Score**: The squared error score is

$$\mathrm{s}(P, x) = (x - \mu_P)^2 \,.$$

   where μ_P is the mean of the predictive distribution P.

6. **Dawid–Sebastiani Score**: The Dawid–Sebastiani score is

$$\mathrm{s}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 + 2\log \sigma_P \,,$$

   where σ_P^2 is the variance of the predictive distribution P.

The scoring rules defined above are used to evaluate the performance of the different models fitted to the data in Section 5.8.2.

Chapter 5

Case study: Association football

A game of football can be viewed as a dynamical system that generates sequences of spatio-temporal events. Analysing those event sequences is directly relevant in the development of game strategies, as well as team and player performance evaluation. However, the analysis of football data is mathematically challenging due to the continuous interaction between the players within and across the two teams. We recognised that event sequences in football can be conveniently represented using marked spatio-temporal point processes and apply the flexible modelling framework developed in this thesis with the goal of describing the game dynamics.

We find the excitation framework of the model proposed in Section 3.2 appropriate for event sequences in football, as any event in the sequence is likely to be triggered by one or more of the previous events. For example, following a corner kick, the next event is almost surely one among a shot on goal, a defensive clearance or a claim by the keeper. In that sense, the corner kick excites the occurrence chance of those three event types in the immediate future.

The proposed model in Section 3.2 captures not only the magnitudes of all cross-excitations between the various event types but also the rate at which these excitations decay over time. The modelling framework along with

second	minute	team_id	player_id	type	outcome	x	y	end_x	end_y
0	0	1	68312	Pass	Successful	49.1	51.0	52.5	44.8
2	0	1	14036	Pass	Successful	52.2	44.5	36.7	60.6
3	0	1	79050	Pass	Successful	36.7	60.6	24.9	39.1
5	0	1	14107	Pass	Unsuccessful	25.0	37.9	97.0	22.9
11	0	2	73379	Tackle	Successful	1.9	73.7	1.9	73.7
15	0	2	73379	Pass	Successful	5.5	65.3	20.9	21.5
17	0	2	6292	Pass	Successful	20.9	21.5	29.0	38.5
19	0	2	26820	Foul	Successful	25.8	37.4	25.8	37.4

Table 5.1: A snapshot of the dataset showing a sequence of events with the relevant attributes.

its parameters can provide valuable insight into the underlying dynamics of the game for the coaching staff and inform strategic decision making. By incorporating covariates such as team information in a direct way as proposed in Section 3.5.1, we are also able to capture the relative abilities of the teams. The efficient simulation framework developed in Section 4.4.2 can be used to obtain instantaneous predictions of goal probabilities, game outcomes or other quantities of interest such as possession ratio or passing accuracy etc. These predictions could enhance, among other things, the viewing experience of televised games.

In this chapter, we describe the football event data and explore the different kinds of analyses such data can be used for. We then provide details on the pre-processing steps used to prepare the dataset for modelling. We define the modelling task and then specify the baseline and excitation-based models employed in this case study. We develop an approach based on association rules to reduce model complexity, before discussing the Bayesian inference for all the fitted models. We then present the results from two approaches for evaluating the accuracy of the Bayesian models. Finally, we provide detailed descriptions for the estimated parameters, and discuss the deep insights they provide about football.

team_id	team name	team_id	team name
1	Arsenal	11	Manchester United
2	Aston Villa	12	Newcastle United
3	Cardiff City	13	Norwich City
4	Chelsea	14	Southampton
5	Crystal Palace	15	Stoke City
6	Everton	16	Sunderland
7	Fulham	17	Swansea City
8	Hull City	18	Tottenham Hotspur
9	Liverpool	19	West Bromwich Albion
10	Manchester City	20	West Ham United

Table 5.2: List of teams competing in the 2013/14 season of the English Premier League.

5.1 Data description

We are provided with event data from all English Premier League games for the 2013/14 and 2014/15 seasons consisting of a record of all touch-ball events within a game. A touch-ball event is an event where a player has acted on the ball by touching it with some part of their body. In total we have about 1.1 million events recorded over the two seasons and a snapshot of the data with the attributes that are relevant to this thesis is provided in Table 5.1. Each season of the league is contested by a total of 20 teams. For example, Table 5.2 gives the list of teams for the 2013/14 season. The league follows a round-robin tournament scheduling, where each team plays every other team at their home and away venues, which equals a total of 760 games over the two seasons.

Each game consists of two halves that are separated by an interruption of approximately 15 minutes. We shall refer to each uninterrupted game half as a game period henceforth. For each event, we have the event type, time-stamp, x,y co-ordinates of its location in the playing field, team and player ids, game period, event outcome (successful/unsuccessful) and the end x,y

event type	frequency	event type	frequency
Pass	747398	SavedShot	9967
BallRecovery	76877	CornerAwarded	8184
Clearance	50729	MissedShots	7822
Tackle	29267	OffsidePass	3057
TakeOn	28224	Claim	2396
BallTouch	27290	Goal	2027
Aerial	26690	Punch	767
Interception	21989	GoodSkill	428
Dispossessed	17590	ShotOnPost	375
Foul	16871	Smother	259
KeeperPickup	10296	CrossNotClaimed	151

Table 5.3: Frequencies of the 22 distinct event types in the dataset.

co-ordinates if the event is a Pass. Table 5.3 gives the frequency of each of the 22 distinct in-game event types recorded in the dataset. The data was provided by Stratagem Technologies based in London, UK.

5.2 Data exploration

We explore the data through a series of visualisations that highlight the potential spatio-temporal data have to capture what happens in a football game. The visualisations also provide an idea about the different kind of analyses such data can be used for.

Ball tracking

A typical way to visualise spatio-temporal data is to trace the location of an object over time. Figure 5.1 shows the trajectory of the ball during an attacking move that lead to a goal in the 18th minute of the game between Arsenal and Norwich City on October 19, 2013. The goal was scored by Jack Wilshere for Arsenal and was voted as the best goal of the season in the English Premier League for the 2013/14 season.

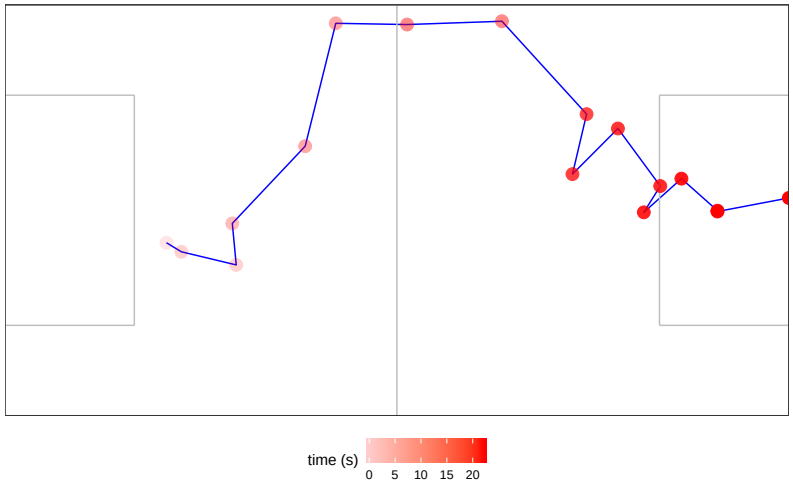


Figure 5.1: Visualising the sequence of events leading up to the goal scored by Jack Wilshere for Arsenal against Norwich City, voted as the Goal of the season (2013/14).

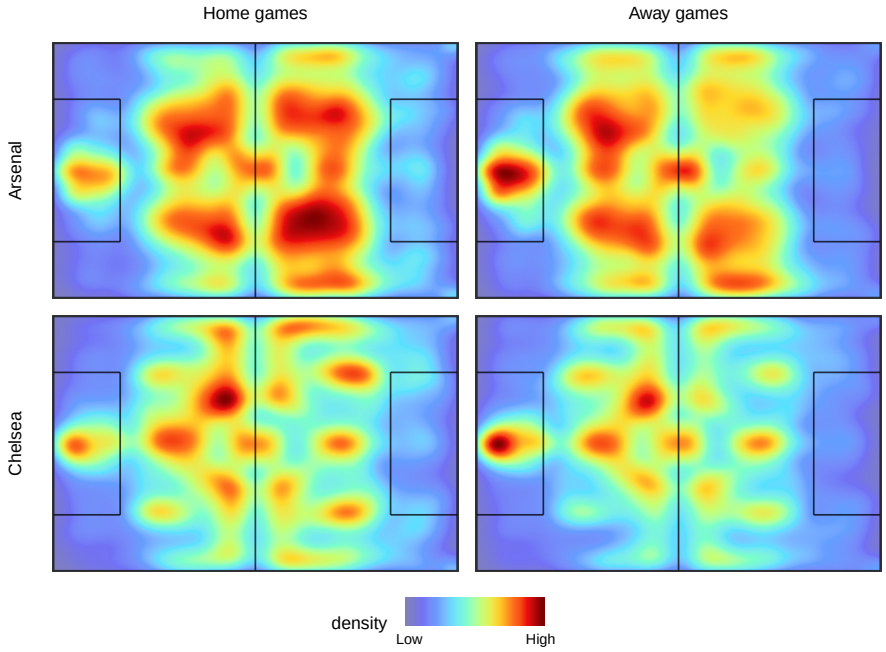


Figure 5.2: Heat map showing the density of ball-touches for Arsenal and Chelsea in their home and away games in the 2013/14 season. In all heat maps the team is attacking to the right, i.e. the opposition goal is to the right.

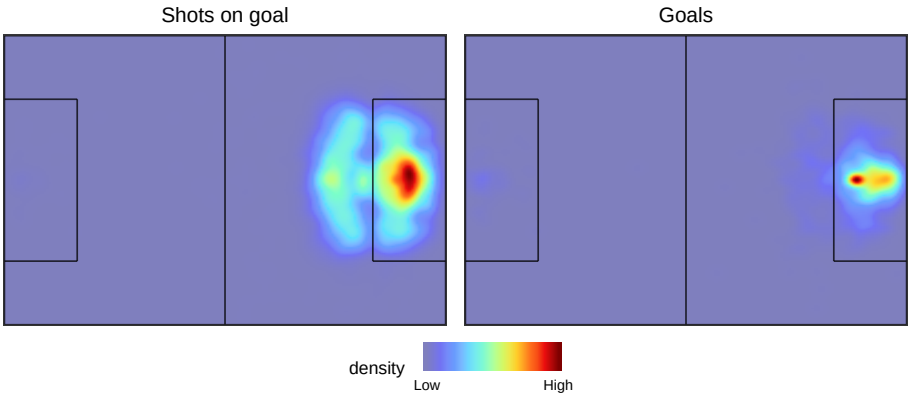


Figure 5.3: Heat map showing the density of all shots attempted on goal (left) vs goals (right) across all teams in the 2013/14 season.

Heat map I: Home advantage

It is a well-known fact in football that teams are more attacking and tend to dominate the opponent in games played at their home venue. We visualise the home advantage phenomenon by comparing the heat maps of ball-touches for Arsenal and Chelsea between their home and away games of the 2013/14 season. Figure 5.2 has the team attacking to the right and we see how the heat maps for the home games for both teams are shifted to the right, i.e. towards the opponents' goal. Using such heat maps we are also able to visualise the difference in the playing styles and formations between the teams.

Heat map II: Shots on Goal

Another interesting aspect to explore is the spatial distribution of the shots attempted on goal. The heat map on the left in Figure 5.3 shows the density of all shots attempted on goal, while the one on the right are the shots that resulted in a goal, from the 2013/14 season. We observe that even though there are a large number of long range shots attempted, the vast majority of them do not result in a goal.

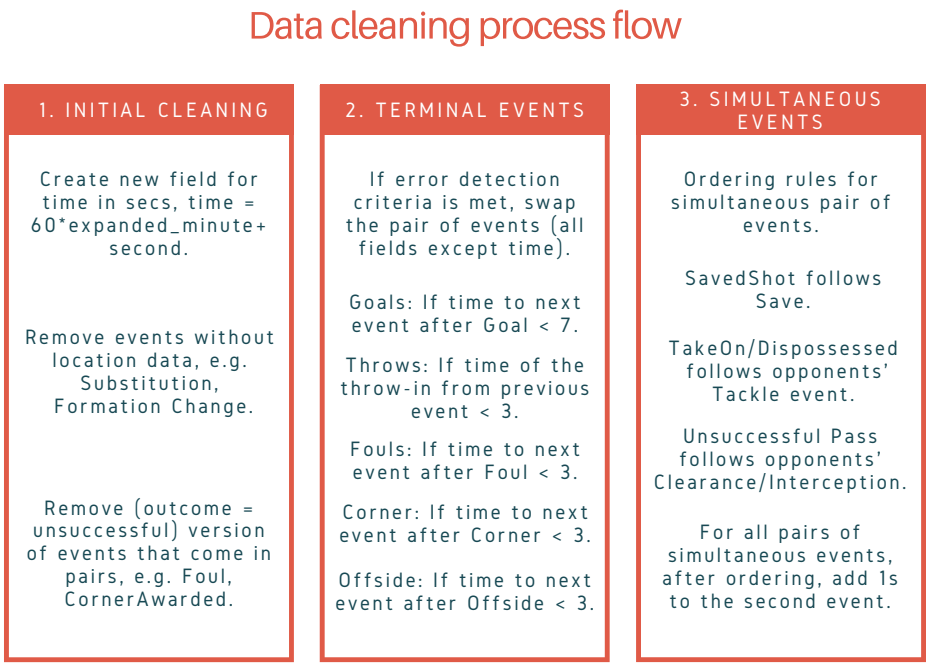


Figure 5.4: Data cleaning workflow showing the steps involved in the three stage process to prepare the dataset for modelling.

5.3 Data pre-processing

In this section, we provide details on the pre-processing treatment that was applied to the raw data to prepare it for modelling.

5.3.1 Cleaning

As the data gathering method was manual and happening at high frequency, erroneous records are understandably quite common. The most critical kind of errors are those where the occurrence order of events has been mixed up. Such errors can result in misleading inferences when studying the causal dependence between events. It is essential, therefore, to develop a systematic way to detect such errors and fix them. Figure 5.4 outlines the three stage procedure applied as part of the data cleaning.

time	team_id	type	outcome
68	1	Clearance	Successful
69	1	CornerAwarded	Unsuccessful
69	2	CornerAwarded	Successful
82	2	Pass	Successful

time	team_id	type	outcome
68	1	Clearance	Successful
69	2	CornerAwarded	Successful
82	2	Pass	Successful

Table 5.4: (Top) Original event sequence with a pair of records for a single CornerAwarded event. (Bottom) Event sequence after removing the outcome = Unsuccessful version of the CornerAwarded event.

Initial cleaning

As a first step, we create a new attribute for the total elapsed time in each game period using the separate minute and second attributes available in the data. We then discard the records of events that are either redundant or unusable because they are missing location data or cannot be classified as touch-ball events. Such records are, for example, player substitutions and formation changes.

We also remove the redundant records of events that come in pairs, such as fouls and corners. These events are recorded for both teams, one for the team successful for receiving the event and one for the opposite team for conceding it. We only retain the record corresponding to the successful version of the event as illustrated in Table 5.4.

Terminal events

Events that result in the ball going out of play are typically followed by a period of inactivity where no events happen. For example, goals and fouls are events that fall into this category of terminal events and we verify the time to the next event after all terminal events. The aim is to detect events

time	team_id	type	outcome
2455	2	CornerAwarded	Successful
2477	2	Pass	Successful
2479	2	Goal	Successful
2482	1	CrossNotClaimed	Successful
2529	1	Pass	Successful

time	team_id	type	outcome
2455	2	CornerAwarded	Successful
2477	2	Pass	Successful
2479	1	CrossNotClaimed	Successful
2482	2	Goal	Successful
2529	1	Pass	Successful

Table 5.5: (Top) Original event sequence with the erroneous CrossNotClaimed event following a Goal event. (Bottom) Event sequence after swapping the Goal and CrossNotClaimed events.

that have been incorrectly recorded to occur shortly after a terminal event. If the detection criteria is met, the pair of events are swapped, i.e., all attributes except time are exchanged between the two events. The window of time for error detection varies according to the type of terminal event as given in Figure 5.4.

We detected a total of 675 pairs of events through the above process across all terminal event types which equals approximately 0.07% of total records. Table 5.5 shows an instance of a CrossNotClaimed event that is incorrectly recorded to happen three seconds after a Goal event. In the corrected version, we have a coherent sequence of events where a corner kick is followed by the goal keeper failing to claim the cross, which results in a goal being scored shortly after.

Simultaneous events

We also have to deal with cases of simultaneous events where a pair of events have been recorded with the same occurrence time. First, we verify

time	team_id	type	outcome
3325	2	Pass	Successful
3328	1	Save	Successful
3328	2	SavedShot	Successful
3332	1	KeeperPickup	Successful

time	team_id	type	outcome
3325	2	Pass	Successful
3328	2	SavedShot	Successful
3329	1	Save	Successful
3333	1	KeeperPickup	Successful

Table 5.6: (Top) Original event sequence with incorrect ordering of Save and SavedShot events. (Bottom) Event sequence after swapping the Save and SavedShot events.

the occurrence order of the event pairs and ensure we avoid any impossible sequences, for example, a Save event by the goalkeeper is followed by a SavedShot event. We compile a list of ordering rules and swap the order of events where necessary. Finally, considering that we use point processes to model these event sequences, where the probability of simultaneous events is zero, we push the occurrence times of all subsequent events after the first event belonging to the simultaneous pair by 1 second.

In this way, we re-ordered a total of 12730 pairs of events, which equals approximately 1.2% of total records. Table 5.6 illustrates how an instance of a SavedShot event following a Save event is dealt with.

5.3.2 Wrangling

Wrangling involves restructuring the data into a desired format for the modelling task.

type	outcome	action	frequency
Clearance	Successful	Clear	50679
Clearance	Unsuccessful	Clear	47
Punch	Successful	Clear	767
TakeOn	Successful	Dribble	13752
Foul	Successful	Foul	16871
Goal	Successful	Goal	2027
Claim	Successful	Keeper	2262
KeeperPickup	Successful	Keeper	11313
BallTouch	Successful	Lose	10291
BallTouch	Unsuccessful	Lose	16999
Claim	Unsuccessful	Lose	134
CrossNotClaimed	Successful	Lose	151
Dispossessed	Successful	Lose	17590
Tackle	Unsuccessful	Lose	6864
TakeOn	Unsuccessful	Lose	14472
CornerAwarded	Successful	Out_Corner	8184
Out_GK	Successful	Out_GK	13026
Out_Throw	Successful	Out_Throw	34505
OffsidePass	Offside	Pass_O	3057
Pass	Successful	Pass_S	577471
Pass	Unsuccessful	Pass_U	169924
Save	Successful	Save	9835
MissedShots	Successful	Shot	7822
SavedShot	Successful	Shot	9967
ShotOnPost	Successful	Shot	375
Aerial	Successful	Win	26690
Interception	Successful	Win	21986
Smother	Successful	Win	259
Tackle	Successful	Win	22403

Table 5.7: Grouping of event types to actions including the frequency of their observations in the dataset.

m	mark label	count	m	mark label	count
1	Home_Win	35608	16	Away_Win	35730
2	Home_Dribble	7103	17	Away_Dribble	6649
3	Home_Pass_S	300320	18	Away_Pass_S	277151
4	Home_Pass_U	85832	19	Away_Pass_U	84092
5	Home_Shot	10161	20	Away_Shot	8003
6	Home_Keeper	6496	21	Away_Keeper	7079
7	Home_Save	4338	22	Away_Save	5497
8	Home_Clear	23528	23	Away_Clear	27965
9	Home_Lose	33296	24	Away_Lose	33205
10	Home_Goal	1157	25	Away_Goal	870
11	Home_Foul	8580	26	Away_Foul	8291
12	Home_Out_Throw	17791	27	Away_Out_Throw	16714
13	Home_Out_GK	5936	28	Away_Out_GK	7090
14	Home_Out_Corner	4627	29	Away_Out_Corner	3557
15	Home_Pass_O	1570	30	Away_Pass_O	1487

Table 5.8: Encoding of marks along with their labels and frequencies in the dataset.

Event grouping

The event outcome attribute is dependent on the event type attribute and hence, we combine the two into what we refer to as actions. At the same step, we also group similar event types together to keep the number of distinct actions to a minimum. This also ensures that we have sufficient observations of each action. Table 5.7 provides details on the grouping of event types into a total of 15 distinct actions. The list of in-play actions includes, Win, Dribble, Successful Pass (Pass_S), Unsuccessful Pass (Pass_U), Shot, Keeper, Save, Clear and Lose events. Terminal actions that result in the ball going out-of-play include, Goal, Foul, Out_Throw, Out_GK, Out_Corner and Offside Pass (Pass_O).

Furthermore, as the same set of actions are tracked for both the home and away teams, we append the string (Home/Away) as a prefix to the action to distinguish between the events of the two teams playing the game. Hence,

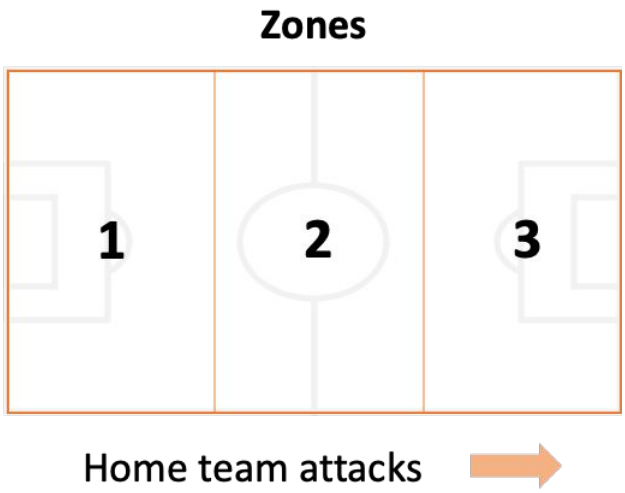


Figure 5.5: Mapping from event location in (x,y) coordinates to zones.

we have a total of $M = 30$ distinct actions in total which form the set of discrete marks in the marked point process definition in Section 2.6. Table 5.8 provides details on the encoding of marks along with their labels and frequencies in the dataset.

Zones for location

The raw data contains location information given by the (x,y) coordinates of the event. To aid with the grouping of events by location we construct a mapping of the (x,y) coordinates into zones, by dividing the length of the playing field equally into three regions. The zones and their corresponding labels are shown in Figure 5.5. For example, zone 1 is the region where the home team defends their goal. The boundaries are constructed with the expectation that the level of control a team has on the game is influenced by the third of the playing field the ball is at. For example, the home team is generally in more control of the game in zone 1, i.e, they're expected to retain possession of the ball more successfully in zone 1 as compared to say, zone 3.

i	id	period	team_id	time (t_i)	zone (z_i)	mark (m_i)
1	101	1	1	0	2	18
2	101	1	1	1	2	19
3	101	1	2	3	1	8
4	101	1	1	6	3	16
5	101	1	1	8	3	18
6	101	1	1	15	2	18
7	101	1	1	16	1	19
8	101	1	2	19	1	12

Table 5.9: Snapshot of the final dataset prepared for modelling.

5.4 Problem definition

A snapshot of the dataset prepared for the modelling task is provided in Table 5.9. The id is a unique identifier for each uninterrupted game period in the dataset and the sequence of events within each game period is modelled as single process.

We can now define the modelling task as follows:

Definition 5.1 *Model the sequence of events in a single period of a football game as a marked spatio-temporal point process, where each event, indexed by $i = 1, ..., n$, consists of the following components,*

1. *time of occurrence t_i,*
2. *zone z_i, and*
3. *mark m_i.*

The home and away team information for each game is assumed to be known and the first event (t_1, z_1, m_1) in each game period is considered to be deterministic and therefore, not modelled.

5.4.1 Likelihood

To specify the likelihood associated with this modelling task, denote the sequences of event times by $\boldsymbol{t} = \{t_i : t_i \in \mathbb{R} \text{ and } t_i > t_{i-1}\}$, locations by $\boldsymbol{z} = \{z_i : z_i \in 1, \ldots, Z\}$ and marks by $\boldsymbol{m} = \{m_i : m_i \in 1, \ldots, M\}\ \forall\ i = 1, \ldots, n$, where n is the number of events in a single game period and Z, M are the number of discrete locations and marks respectively. The likelihood for the sequence of events within a game period based on a model with parameter vector $\boldsymbol{\Theta}$ is

$$\mathcal{L}(\boldsymbol{\Theta}) = p(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{m} \mid \boldsymbol{\Theta})\,. \tag{5.1}$$

All game periods are modelled independently, and therefore, the likelihood for a collection of game periods can be calculated by simply taking a product of the individual likelihoods.

5.4.2 Training data

All models are trained on the first 20 games (40 game periods) of the 2013/14 season. The training data consists of a total of 27,660 events over the 40 game periods and each of the 20 teams plays two games, one each at their home and away venues. Table 5.10 gives the zone-wise event frequencies in the training data used for the modelling experiment.

5.5 Models employed

In this section, we provide details on the specification for all the models that are fitted to the data. Broadly we divide the models into two categories; those with Hawkes-like excitation effects and ones without excitation that are used as baseline models.

mark		zone		
m	label	1	2	3
1	Home_Win	236	257	41
2	Home_Dribble	17	96	96
3	Home_Pass_S	1699	4633	1658
4	Home_Pass_U	541	825	725
5	Home_Shot	0	0	292
6	Home_Keeper	155	0	0
7	Home_Save	106	0	0
8	Home_Clear	557	141	32
9	Home_Lose	122	287	368
10	Home_Goal	0	0	22
11	Home_Foul	62	126	64
12	Home_Out_Throw	97	184	163
13	Home_Out_GK	149	0	0
14	Home_Out_Corner	0	0	112
15	Home_Pass_O	7	19	20
16	Away_Win	25	204	301
17	Away_Dribble	76	65	19
18	Away_Pass_S	1427	4390	2030
19	Away_Pass_U	542	811	702
20	Away_Shot	193	2	0
21	Away_Keeper	0	0	192
22	Away_Save	0	0	149
23	Away_Clear	27	124	660
24	Away_Lose	323	349	142
25	Away_Goal	20	0	0
26	Away_Foul	46	112	69
27	Away_Out_Throw	143	173	110
28	Away_Out_GK	0	0	220
29	Away_Out_Corner	76	0	0
30	Away_Pass_O	13	11	5

Table 5.10: Zone-wise event frequencies in the training data used for the modelling experiment.

All models, except for the homogeneous Poisson model in Section 5.5.1, are specified based on the factorisation of the likelihood in expression (5.1). As seen in expression (3.9), the likelihood of a marked spatio-temporal point process can be factorised as

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} h(z_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\eta}) \prod_{i=1}^{n} f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}), \quad (5.2)$$

where g, h and f are the conditional probability distribution functions for the times, the locations and the marks respectively, and $\boldsymbol{\Theta} = \{\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\theta}\}$ are the corresponding unknown parameter vectors. Therefore, a valid model results by specifying the distribution functions g, h and f individually.

The primary goal in the modelling experiment is to compare how the different models perform in capturing the dependence between the various event types. Keeping this in mind, we opted to use the same specification for the occurrence times g and locations h, and only vary the model for the marks f across all the fitted models. The only exception is the baseline homogeneous Poisson model in Section 5.5.1, which uses a joint specification for the times, locations and marks.

The probability density function for the occurrence times is set to

$$\begin{aligned} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) &= p(t_i - t_{i-1} \mid m_{i-1}, \boldsymbol{a}, \boldsymbol{b}) \\ t_i - t_{i-1} \mid m_{i-1}, \boldsymbol{a}, \boldsymbol{b} &\sim \textbf{Gamma}[a_{m_{i-1}}, b_{m_{i-1}}]. \end{aligned} \quad (5.3)$$

The time to next event is modelled using a gamma distribution with shape and rate parameters that are specific to the mark of the last observed event. With the specification in expression (5.3), we wish to capture the differences in the expected time to the next event across the different event types. For example, we expect a significantly shorter delay before the next event following an in-play event like a Pass compared to that of an out-of-play event like a Foul. Section 5.3.2 provided details on the definition of the event

groups. Even within the group of out-of-play events, for example, we expect a shorter delay following a Throw-in as compared to a Goal event.

The probability mass function for the locations is set to

$$\begin{aligned} h(z_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\eta}) &= p(z_i \mid z_{i-1}, m_{i-1}, \boldsymbol{\eta}) \\ &= \eta_{(z_{i-1}, m_{i-1}) \to z_i}, \end{aligned} \tag{5.4}$$

where $\eta_{(z_{i-1}, m_{i-1}) \to z_i}$ is the probability of transitioning into location z_i given the location z_{i-1} and the mark m_{i-1} of the last observed event. Expression (5.4) models the sequence of locations as a discrete time first order Markov chain (see, for example, Norris, 1997) with a transition probability matrix $\boldsymbol{\eta}$. The current state of the Markov chain is defined by the combination of the location and the mark of the last observed event, and the probability of transitioning into the next location depends only on the current state. The state space of the Markov chain is given by the Cartesian product $\{1, \ldots, Z\} \times \{1, \ldots, M\}$. Note that the transition probability matrix $\boldsymbol{\eta}$ is not a square matrix as we only model the next location using the function h and the model for the marks is specified separately by the function f.

5.5.1 Baseline models

First, we specify the two baseline methods that can be considered as the competition for the models with excitation effects specified in Section 5.5.2. The intention is to compare the baseline models against the excitation-based models using the model evaluation techniques detailed in Section 4.4.

Homogeneous Poisson process

The simplest non-trivial model for marked spatio-temporal data is the homogeneous Poisson process model. We fit individual homogeneous Poisson processes for each mark in each location in the training data. This corresponds to a basic featureless predictor that assumes no signal in the data.

The log-likelihood for the marked spatio-temporal homogeneous Poisson process model (Daley and Vere-Jones, 2003) is

$$\log(\mathcal{L}) = \sum_{m=1}^{M} \sum_{z=1}^{Z} \left[N_{m,z} \log(r_{m,z}) - T\, r_{m,z} \right] ,$$

where $r_{m,z}$ is the Poisson rate parameter and $N_{m,z}$ is the number of event occurrences for mark m in location z respectively over a total duration of time T.

Markov chain

As a second baseline, we model the sequence of marks using a Markov chain, similar to the model for the sequence of locations in expression (5.4). The probability mass function for the marks is set to

$$\begin{aligned} f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) &= p(m_i \mid z_i, m_{i-1}, \boldsymbol{\theta}) \\ &= \theta_{(z_i, m_{i-1}) \to m_i} , \end{aligned} \tag{5.5}$$

where $\theta_{(z_i, m_{i-1}) \to m_i}$ is the probability of the event mark m_i given the event location z_i and mark m_{i-1} of the last observed event.

Unlike, the homogeneous Poisson model, the Markov chain in expression (5.5) has memory of the first order and is a popular model for a wide range of real-world processes. We use the specifications for the times and locations in expressions (5.3) and (5.4) to complete the model specification for the Markov chain based baseline.

5.5.2 Excitation based models

To model the sequence of marks, the models in this section use extensions of the probability mass function for the marks in expression (3.4) derived from the marked Hawkes process intensity. Recall that the specification in expression (3.4) captures all cross-excitations between the various marks as well as the rates at which these excitations decay over time. Those were the

crucial features of the marked Hawkes process model we wished to retain in the modelling framework we introduced. The following models all use the same specifications for the times and locations in expressions (5.3) and (5.4) and therefore, we restrict the discussion to their specific model for the marks.

Scalar beta

The Scalar beta model is the basic specification for the marks in expression (3.4), where the excitation decay rate β is a scalar. The probability mass function for the marks from expression (3.4) is

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta(t_i - t_j)}} \,. \tag{5.6}$$

Vector beta

The probability mass function for the marks in the Vector beta model is

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta_{m_j}(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta_{m_j}(t_i - t_j)}} \,, \tag{5.7}$$

where β_{m_j} is the exponential decay rate of the excitation caused by an event of mark m_j. By allowing the decay rates to depend on the mark of the event causing the excitation, the specification in expression (5.7) is more flexible than the Scalar beta model in expression (5.6).

Matrix beta

The Matrix beta model is one of the key model extensions proposed in Section 3.5.2, where the decay rate β depends on the pair of marks involved in the excitation. Additionally, we also allow the decay rates, the background mark probability δ and event conversion rates γ to vary across the different

locations as

$$f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i|z_i} + \sum_{t_j<t_i} \mathrm{e}^{\alpha-\beta_{m_j\to m_i|z_i}(t_i-t_j)}\gamma_{m_j\to m_i|z_i}}{\sum_{m=1}^{M}\left[\delta_{m|z_i} + \sum_{t_j<t_i} \mathrm{e}^{\alpha-\beta_{m_j\to m|z_i}(t_i-t_j)}\gamma_{m_j\to m|z_i}\right]}, \quad (5.8)$$

where $\beta_{m_j\to m_i|z_i}$ is the decay rate of the excitation caused by an event of mark m_j on an event of mark m_i in the location z_i. The specification in expression (5.8) offers maximum flexibility for the excitation effects to vary across different mark pairs and can model dependence between events over arbitrary lengths of time. Such flexibility is essential to account for scenarios like a Corner event exciting a Pass event in the immediate future and a Shot event later on, i.e. $\beta_{\mathsf{Corner}\to\mathsf{Pass}} > \beta_{\mathsf{Corner}\to\mathsf{Shot}}$.

The dependence on location of the decay rates, the background mark probabilities and the event conversion rates in expression (5.8) allows us to capture effects like how a team is more likely to make more passes and retain possession of the ball in the defensive third, while attempting shots on goal in the attacking third, i.e., for the home team, we have $\gamma_{\mathsf{Pass}\to\mathsf{Pass}|1} > \gamma_{\mathsf{Pass}\to\mathsf{Pass}|3}$ and $\gamma_{\mathsf{Pass}\to\mathsf{Shot}|3} > \gamma_{\mathsf{Pass}\to\mathsf{Shot}|1}$.

Including team information in the Matrix beta model

Another key model extension discussed in Section 3.5.1 is the incorporation of covariates in the conversion rate parameters γ using a baseline-category logit specification. In the model for football data, team information is incorporated as

$$\log\left(\frac{\gamma_{m_j\to m|z}(h)}{\gamma_{m_j\to M|z}(h)}\right) = \varphi_{m_j\to m|z} + \omega_{h,m} \qquad \forall\, m \in 1,\ldots,M-1\,, \quad (5.9)$$

where φ is the location dependent baseline conversion parameter and h is the team in possession of the ball attempting the event conversion. The parameter ω is then interpreted as the relative ability of a team to complete a conversion to an event of mark m.

We fit the Matrix beta model to the data as specified in expression (5.8) both with and without the inclusion of team information via expression (5.9). However, we do not include team information while fitting the simpler Scalar beta model in expression (5.6) or the Vector beta model in expression (5.7).

5.6 Dealing with model complexity

The model for the marks specified in the Matrix beta model in expression (5.8) introduces $\mathcal{O}(M^2)$ number of parameters into the model. For a total of $M = 30$ event types and 3 locations, we have $30 \times 30 \times 3 = 2700$ decay rate parameters and $30 \times 29 \times 3 = 2610$ conversion rate parameters, bringing the total number of parameters up to 5950. Estimating the large number of parameters from this model is not inherently problematic, but the limited availability of computational resources render the estimation a challenging task. There is also the added risk of over-fitting that we have to guard against. In this section, we propose an algorithm to deal with such model complexity by reducing the number of estimated parameters using association rule learning.

In the Matrix beta model in expression (5.8), the decay rate parameters $\boldsymbol{\beta}$ and conversion rate parameters $\boldsymbol{\gamma}$ capture the duration and magnitude of the excitation effects between all pairs of event types. However, it is reasonable to assume that the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are sparse, because the excitation effects between all event pairs are not equally significant. To be precise, we expect most elements of the $\boldsymbol{\beta}$ matrix to be infinite, meaning the corresponding excitations decay almost instantaneously. For the $\boldsymbol{\gamma}$ matrix, we expect most its values to be zero, meaning the corresponding event conversions have probability zero. For example, a successful Pass event by one team cannot significantly excite a Pass event for the opposite team, as this would

make the commonplace occurrence of a string of passes by a single team very unlikely. If we are able to identify the most significant pairs of event interactions, we can thereby limit the number of elements within the matrices β and γ that we need to estimate.

5.6.1 Association rule learning

Association rule learning is a method for discovering strong relationships between variables in large databases (see, for example, Agrawal et al., 1993). For example, the association rule Bread $\Rightarrow$ Butter identified from a supermarket sales database would indicate that if a customer buys bread, they are also likely to buy butter. The objective of association rule learning is to identify rules that are interesting based on some measure of significance.

5.6.2 Definition for event sequences

Inspired by the original definition in Agrawal et al. (1993, Section 2), we define the problem of association rule learning in the context of event sequences as

Definition 5.2

- *Let* $A = \{1, \ldots, M\}$ *be the set of M distinct event types.*
- *Let* $B = \{b_{s,n}\}$, *where* $b_{s,n} \in A$ *for* $s \in \{1, \ldots, S\}$ *and* $n \in \{1, \ldots, N_s\}$, *be the training data consisting of S event sequences with* N_s *number of observed events in the sequence s.*
- *Construct a database of subsequences* $D = \{d_1, \ldots, d_C\}$, *where* $C = \sum_1^S N_s$, *such that each event b in B has a corresponding subsequence of length* $W + 1$ *in D, made up of b and the W events preceding b.*
- *Each subsequence in D is denoted by* $d_i = \{x_{i,1}, \ldots, x_{i,W}, y_i\}$, *where* $x_{i,j}, y_i \in B$ *for* $i \in \{1, \ldots, C\}$ *and* $j \in \{1, \ldots, W\}$. *We call* $\{x_{i,1}, \ldots, x_{i,W}\}$ *as the*

transient events of the subsequence before the terminal event y_i. Depending on W, the elements of the subsequence corresponding to the initial events of a sequence can be empty, because they have shorter histories.

- *Given a set of event types A and a database of subsequences D, a rule is defined as an implication of the form: $x \Rightarrow y$, where $x, y \in A$. The association rule has the interpretation that the event type x is likely to be a transient event in subsequences terminating with event type y.*

In other words, the rule $x \Rightarrow y$, would indicate that the event type x excites the occurrence chance of an event with type y.

5.6.3 Measures of significance

To identify interesting association rules, we place constraints on two measures of significance (Brin et al., 1997), namely support and lift.

Support

The support of x with respect to a rule $x \Rightarrow y$ and a database D is defined as the proportion of subsequences d in the database which contain x as a transient event,

$$P(x) = \frac{|\{d \in D; x \in \mathsf{trans}(d)\}|}{|D|}, \tag{5.10}$$

where $|\cdot|$ denotes the cardinality of a set and $\mathsf{trans}(d)$ is the set of transient events in the subsequence d. Similarly, the support of y with respect to a rule $x \Rightarrow y$ is defined as the proportion of subsequences d which terminate with y,

$$P(y) = \frac{|\{d \in D; y \in \mathsf{term}(d)\}|}{|D|}, \tag{5.11}$$

where $\mathsf{term}(d)$ is the terminal event in the subsequence d.

	Home_Win	Home_Dribble	Home_Pass_S	Home_Pass_U
Home_Win	0.0015	0.0027	0.0663	0.0124
Home_Dribble	0.0006	0.0008	0.0158	0.0030
Home_Pass_S	0.0111	0.0099	0.5925	0.0962
Home_Pass_U	0.0163	0.0026	0.1036	0.0289

Table 5.11: Support $P(x \cap y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y.

	Home_Win	Home_Dribble	Home_Pass_S	Home_Pass_U
Home_Win	0.4141	2.2669	0.9793	0.9766
Home_Dribble	0.7176	2.7990	0.9588	0.9698
Home_Pass_S	0.3845	1.0141	1.0879	0.9450
Home_Pass_U	2.3031	1.0860	0.7782	1.1609

Table 5.12: $\mathsf{lift}(x \Rightarrow y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y.

The support of a rule $x \Rightarrow y$ is defined as, the proportion of subsequences d which contain x as a transient event and terminate in y,

$$P(x \cap y) = \frac{|\{d \in D; x \in \mathsf{trans}(d); y \in \mathsf{term}(d)\}|}{|D|}. \tag{5.12}$$

Table 5.11 gives the support $P(x \cap y)$ for selected event pairs in the training data.

Lift

The lift of a rule $x \Rightarrow y$ is defined as

$$\mathsf{lift}(x \Rightarrow y) = \frac{P(x \cap y)}{P(x) \cdot P(y)}. \tag{5.13}$$

If the lift of a rule equals 1, it would indicate that the occurrence of y is independent of that of x. If the rule has $\mathsf{lift} > 1$, then the event x excites the occurrence chance of y and $\mathsf{lift} < 1$ indicates x inhibits the occurrence of y. Table 5.12 gives the $\mathsf{lift}(x \Rightarrow y)$ for selected event pairs in the training data.

We implement the following steps to place constraints on the lift and support measures and identify significant dependence between pairs of events.

- Create a database of subsequences as defined in Definition 5.2, for $W = 5$ and $W = 10$, where W is the number of transient events in each subsequence.
- For each W, calculate lift for all event pairs and retain only those pairs that have lift > 1.
- Set a threshold on the support $P(x \cap y) > \varepsilon$, such that when $\varepsilon = \varepsilon_1$ exactly $N = 50$ event pairs remain, and when $\varepsilon = \varepsilon_2$, $N = 100$ event pairs remain.

In this way, we select the specific elements of the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, corresponding to the identified significant event pairs, for parameter estimation. The elements of the matrices corresponding to the discarded event pairs are fixed, to the value 10^6 in the case of the decay rates $\boldsymbol{\beta}$, and 10^{-6} for the conversion rates $\boldsymbol{\gamma}$. A large value for the decay rate causes the excitation to die out almost instantaneously, and a very small value for the conversion rate makes the event conversion extremely unlikely. The results of evaluating four separate models, that are fitted based on the specific choices of the tuning parameters given above for the length of subsequence window W and the number of identified event pairs N, are discussed in Section 5.8.

5.7 Bayesian inference

In this section, we provide details on the Bayesian inference for the fitted models in the same order as they are presented in Section 5.5. As noted in Remark 3.1, we are able to do parameter estimation for the components of the process, namely, the times, the locations and the marks, separately as the component-wise models in the factorisation in expression (5.2) do not share

any parameters.

Inference for the model for the inter-arrival times in expression (5.3) and the excitation based models for the marks in Section 5.5.2, are carried out via the Bayesian framework proposed in Chapter 4 that uses an HMC algorithm for sampling from the posterior distribution. We present descriptive summaries of the generated parameter samples as well as diagnostics to evaluate the convergence of the sampling algorithm.

The model for the locations in expression (5.4) and the baseline models in Section 5.5.1 have closed-from expressions for their posterior distributions that can be sampled from without the need for an MCMC algorithm. Specifically, in Sections 5.7.2, 5.7.3 and 5.7.4, we exploit the fact that the likelihoods for these models have a conjugate prior and therefore, the posterior distribution can be calculated by simply updating the parameters of the prior distribution using a set of sufficient statistics (Raiffa and Schlaifer, 1961).

5.7.1 Gamma process model for the occurrence times

The inter-arrival times are modelled using a Gamma distribution with parameters specific to the mark of the last observed event as specified in expression (5.3). Recall that this Gamma process model in expression (5.3) is just one component of the complete model specification in expression (5.2). However, we are able to perform inference for each component separately as they do not share any parameters. We use exponential priors for the Gamma process model parameters as specified in expression (4.5), where the hyper-parameters a' and b' are both set to 0.01.

We obtain posterior parameter samples by running three parallel chains of the HMC algorithm, implemented in Stan as discussed in Section 4.2.3. Each chain is initialised with a different starting value and run for a total of 1000 iterations post the warm-up phase. Figure 5.6 is a chain-wise trace plot for

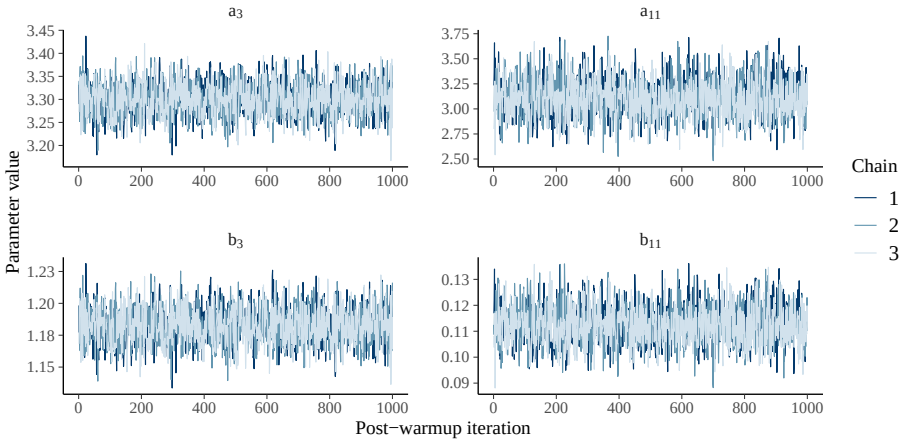


Figure 5.6: Trace plot of the posterior samples for selected parameters across multiple chains from the Gamma process model for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i.

parameter	mean	sd	$\hat{R}$	N_{eff}
a_3	3.30	0.04	1.00	1827.48
a_5	10.24	0.64	1.00	1721.74
a_{10}	64.40	13.16	1.00	1789.70
a_{11}	3.11	0.19	1.01	1581.00
b_3	1.18	0.01	1.00	1940.04
b_5	9.28	0.59	1.00	1771.61
b_{10}	1.25	0.26	1.00	1800.29
b_{11}	0.11	0.01	1.01	1543.40

Table 5.13: Posterior summaries and convergence diagnostics from 3000 posterior samples for selected parameters from the Gamma distribution for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i.

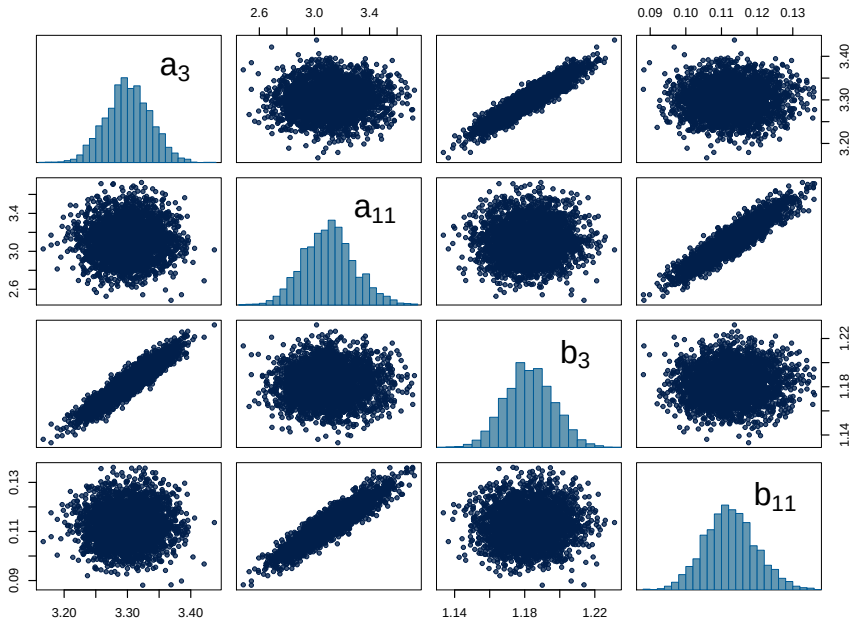


Figure 5.7: Pair-wise correlations with marginals along the diagonal for selected model parameters from the Gamma distribution for inter-arrival times in expression (5.3). a_i and b_i are the shape and rate parameters respectively corresponding to mark i.

four of the parameters, created using the R package bayesplot (Gabry and Mahr, 2019). Table 5.13 gives the descriptive statistics along with the convergence diagnostics. We monitor the convergence of the algorithm using the potential scale reduction factor $\hat{R}$ proposed by Gelman et al. (1992). The potential scale reduction factor $\hat{R}$ measures the ratio of the average variance within each chain to the variance of the aggregated samples across chains. If the algorithm has converged and the chains are at equilibrium, $\hat{R}$ equals 1. From the posterior samples, we confirm that all parameters have $\hat{R} < 1.1$ as recommended in Gelman et al. (1992) as a test for convergence.

Posterior sampling algorithms can also suffer from autocorrelation within a chain, which for example, increases the uncertainty in the estimates of posterior means and variances (Geyer, 2011). The autocorrelation within

Mark label	a_i (shape)	b_i (rate)	a_i/b_i (mean)
Win	2.94	1.68	1.75
Dribble	2.88	1.23	2.35
Pass_S	3.30	1.18	2.79
Pass_U	2.60	1.04	2.51
Shot	10.24	9.28	1.10
Keeper	1.34	0.16	8.61
Save	3.01	0.89	3.39
Clear	2.58	1.17	2.21
Lose	3.31	2.19	1.51
Goal	64.40	1.25	51.62
Foul	3.11	0.11	27.63
Out_Throw	2.83	0.19	14.82
Out_GK	10.08	0.34	29.31
Out_Corner	9.42	0.40	23.79
Pass_O	6.79	0.24	28.40

Table 5.14: Posterior means of the event specific shape and rate parameters of the Gamma distribution for inter-arrival times in expression (5.3) for in-play (top) and out-of-play events (bottom). The column a_i/b_i gives the posterior mean of the expected value of the Gamma distribution.

chains can be measured using the effective sample size statistic, which calculates the effective number of independent samples (see, for example, Gelman et al., 2013, Section 11.5). The computed effective sample sizes (N_{eff} column in Table 5.13) indicate that the sampler returned samples with satisfactory autocorrelation, as the minimum effective sample size is at least 10% of the total sample size. Figure 5.7 is a correlation plot that gives us an idea of the inherent correlations present in the model between the corresponding shape and rate parameters for each mark.

Recall that, the inter-arrival times are modelled with parameters specific to the mark of the last observed event and Table 5.14 gives the posterior parameter means for the mark-specific shape and rate parameters as well as the mean of the Gamma distribution. We do not distinguish between the marks of the home and away teams in this model, which means, for example,

state i		next zone j		
zone	mark label	1	2	3
1	Home_Win	195	38	1
1	Home_Dribble	12	5	0
1	Home_Pass_S	845	797	51
1	Home_Pass_U	75	304	160
1	Home_Shot	0	0	0

Table 5.15: Observed transition counts $y_{i \to j}$ from the first 5 states to each zone in the training data.

the expected time to the next event is the same following a Home_Shot or an Away_Shot event. We observe a clear distinction in the expected time to the next event (mean column in Table 5.14) following an in-play event compared to that of an out-of-play event. This accounts for the typical delays we observe before the game restarts once the ball goes out of play.

5.7.2 Markov chain model for the locations

The probability mass function for the locations specified in expression (5.4), models the locations as a multinomial distribution given the current state (defined by the location and the mark of the last observed event). Similar to the model for the inter-arrival times, the model for the locations in expression (5.4) is another component of the complete model specification in expression (5.2). Once again, we are able to perform inference for this model separately as it does not share any parameters with the other components. Each row of the transition probability matrix $\boldsymbol{\eta}$, corresponding to a single state, is a set of multinomial parameters, one for each location, that add up to 1.

Let $\boldsymbol{y} = \{y_{i \to j}\}$, for $j \in \{1, \ldots, Z\}$, be the observed counts of transitions originating from the state i where $i \in \{1, \ldots, Z\} \times \{1, \ldots, M\}$. Table 5.15 gives the observed transition counts from the first 5 states in the training data.

state i		next zone j		
zone	mark label	1	2	3
1	Home_Win	0.83	0.16	0.01
1	Home_Dribble	0.65	0.30	0.05
1	Home_Pass_S	0.50	0.47	0.03
1	Home_Pass_U	0.14	0.56	0.30
1	Home_Shot	0.33	0.33	0.33

Table 5.16: Posterior means of the multinomial probabilities $\eta_{i\to j}$ for transitions from the first 5 states.

Out of a total of 90 states, 23 are never observed in the dataset, for example, it is nearly impossible for a Home_Shot event to occur in the defensive third (zone = 1) of the home team.

The likelihood of $\boldsymbol{y}_i$ given the multinomial probabilities $\boldsymbol{\eta}_i$ is

$$p(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i) \propto \prod_{j=1}^{Z} \eta_{i\to j}^{y_{i\to j}},$$

where $\sum_{j=1}^{Z} \eta_{i\to j} = 1$. The conjugate prior for the multinomial distribution is the Dirichlet distribution (see, for example, Gelman et al., 2013, Section 3.4),

$$p(\boldsymbol{\eta}_i \mid \mathbf{v}_i) \propto \prod_{j=1}^{Z} \eta_{i\to j}^{v_{i\to j}-1},$$

where $\mathbf{v}_i > 0$ are the hyperparameters. The posterior distribution of $\boldsymbol{\eta}_i$ is therefore a Dirichlet with parameters $\mathbf{v}_i + \boldsymbol{y}_i$.

The hyperparameters $\mathbf{v}_i$ are set to 1 and the resulting posterior means of the parameters $\eta_{i\to j}$ are given in Table 5.16. We use the rdirichlet function from the R package MCMCpack by Martin et al. (2011) to generate samples from the posterior distribution, which are used in the model evaluation performed in Section 5.8.

mark		zone		
m	label	1	2	3
1	Home_Win	0.0035	0.0038	0.0006
2	Home_Dribble	0.0003	0.0014	0.0014
3	Home_Pass_S	0.0251	0.0683	0.0244
4	Home_Pass_U	0.0080	0.0122	0.0107
5	Home_Shot	0.0000	0.0000	0.0043

Table 5.17: Posterior means of the homogeneous Poisson rates $r_{m,z}$, for the first 5 marks in each zone.

5.7.3 Baseline homogeneous Poisson process model

The likelihood for the homogeneous Poisson model for marked spatio temporal data as specified in Section 5.5.1 is

$$\mathcal{L} = \prod_{m=1}^{M} \prod_{z=1}^{Z} r_{m,z}^{N_{m,z}} \exp\left(-T\, r_{m,z}\right) ,$$

where $r_{m,z}$ is the Poisson rate parameter and $N_{m,z}$ is the number of event occurrences for mark m in location z respectively over a total duration of time T. Table 5.10 gives the observed counts $N_{m,z}$ in the training data. The conjugate prior for the Poisson process likelihood is a Gamma distribution

$$p(\boldsymbol{r} \mid \kappa, \tau) \propto \prod_{m=1}^{M} \prod_{z=1}^{Z} r_{m,z}^{\kappa-1} \exp\left(-\tau\, r_{m,z}\right) ,$$

where $\kappa > 0$ and $\tau > 0$ are the hyperparameters for the shape and rate of the Gamma distribution respectively. Therefore, the posterior distribution of $\boldsymbol{r}$ is a Gamma distribution

$$\kappa' = \kappa + N_{m,z} \qquad \tau' = \tau + T\,,$$

where κ' and τ' are the updated hyperparameters. We set the values, $\kappa = 1$ and $\tau = 0$ that correspond to a non-informative prior.

The resulting posterior means of the Poisson rates $r_{m,z}$, for the first 5 marks in each zone, are given in Table 5.17. We use the rgamma function from the

state i		label of next mark j				
mark label	zone	H_Win	H_Dribble	H_Pass_S	H_Pass_U	H_Shot
H_Win	1	0	0	80	25	0
H_Win	2	0	8	138	18	0
H_Win	3	0	4	28	11	4
H_Dribble	1	1	1	8	3	0
H_Dribble	2	0	5	39	11	0

Table 5.18: Transition counts $c_{i \to j}$ from the first 5 states to the first 5 marks in the training data. We abbreviate the prefix Home to H in the mark labels.

R package stats, which implements the method proposed by Ahrens and Dieter (1982), for simulating from a Gamma distribution.

5.7.4 Baseline Markov chain model for the marks

The probability mass function for the marks specified in expression (5.5.1), models the marks as a multinomial distribution given the current state (defined by the current location and the mark of the last observed event). Each row of the transition probability matrix $\boldsymbol{\theta}$, corresponding to a single state, is a set of multinomial parameters, one for each mark, that add up to 1.

Similar to the model for locations in Section 5.7.2, let $\boldsymbol{c} = \{c_{i \to j}\}$, for $j \in \{1, \ldots, M\}$, be the count of observations of the transitions from the state i where $i \in \{1, \ldots, M\} \times \{1, \ldots, Z\}$. Table 5.18 gives the observed counts of transitions from the first 5 states in the training data.

The likelihood of $\boldsymbol{c}$ given the multinomial parameters $\boldsymbol{\theta}$ is

$$p(\boldsymbol{c}_i \mid \boldsymbol{\theta}_i) \propto \prod_{j=1}^{M} \theta_{i \to j}^{c_{i \to j}},$$

where the sum of the probabilities, $\sum_{j=1}^{M} \theta_{i \to j} = 1$. The conjugate prior for the multinomial distribution is the Dirichlet distribution,

$$p(\boldsymbol{\theta}_i \mid \mathbf{u}_i) \propto \prod_{j=1}^{M} \theta_{i \to j}^{u_{i \to j} - 1},$$

state i		label of next mark j				
mark label	zone	H_Win	H_Dribble	H_Pass_S	H_Pass_U	H_Shot
H_Win	1	0.01	0.01	0.74	0.24	0.01
H_Win	2	0.01	0.05	0.82	0.11	0.01
H_Win	3	0.02	0.10	0.56	0.23	0.10
H_Dribble	1	0.11	0.11	0.50	0.22	0.06
H_Dribble	2	0.02	0.10	0.67	0.20	0.02

Table 5.19: Posterior means of the multinomial parameters $\theta_{i\to j}$ corresponding to the first 5 states. We abbreviate the prefix Home to H in the mark labels.

where $\mathbf{u}_i > 0$ are the hyperparameters. The posterior distribution of $\boldsymbol{\theta}_i$ is therefore a Dirichlet with parameters $\mathbf{u}_i + \boldsymbol{c}_i$. We set $\mathbf{u}_i$ to 1 and the resulting posterior means of the parameters $\eta_{i\to j}$ corresponding to the first 5 states are given in Table 5.19.

5.7.5 Excitation based models for the marks

Section 5.5.2 provides details on three different specifications for the excitation based model for the marks, namely the Scalar beta, the Vector beta and the Matrix beta models with the optional inclusion of team information as a covariate. We restrict ourselves to discussing the inference for the Matrix beta model with team information as it is the most comprehensive and the other models are simpler versions of it.

We consider the probability mass function for the marks in the Matrix beta model as specified in expression (5.8), including the baseline logit specification in expression (5.9) to incorporate team information into the conversion rate parameters γ. We also implement the rule-based framework in Section 5.6 for limiting the number of estimated parameters in the matrices $\boldsymbol{\beta}$ and γ. As an example, we discuss the results for a specific choice of the tuning parameters, in which we set the window size $W = 5$ for the number of transient events and identify $N = 100$ event pairs in each of the three zones.

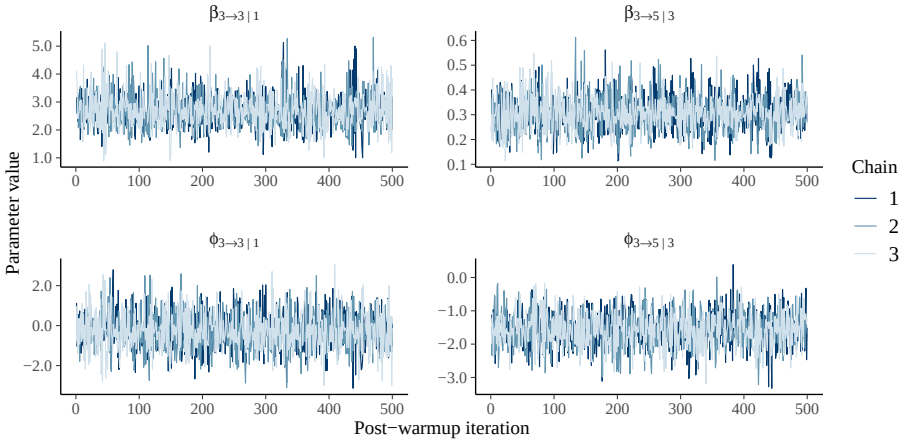


Figure 5.8: Trace plot of the posterior samples for selected parameters across multiple chains from the Matrix beta model for the marks in expression (5.8).

parameter	mean	sd	$\hat{R}$	N_{eff}
$\beta_{3\to3\|1}$	2.70	0.64	1.00	800.84
$\beta_{24\to1\|2}$	1.49	0.08	1.00	1229.42
$\beta_{3\to5\|3}$	0.30	0.07	1.01	1219.00
$\varphi_{3\to3\|1}$	-0.28	0.98	1.00	1104.87
$\varphi_{24\to1\|2}$	3.52	0.35	1.00	1300.19
$\varphi_{3\to5\|3}$	-1.56	0.54	1.00	817.92
$\delta_{10\|3}$	0.02	0.00	1.00	1350.87
α	6.28	0.08	1.01	1042.16

Table 5.20: Posterior summaries and convergence diagnostics from 1500 posterior samples for selected parameters from the Matrix beta model for the marks in expression (5.8).

The prior distributions for the model parameters are as specified in Section 4.1.3, with the following setting for the hyperparameters. The Dirichlet prior on the background mark probabilities $\boldsymbol{\delta}$ has concentration hyperparameters $\delta' = 1$. The exponential prior on the decay rates $\boldsymbol{\beta}$ has a rate hyperparameter $\beta' = 0.1$. The Normal priors on the excitation factor α and the baseline-category logit model parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}$ have hyperparameters $\sigma_\alpha, \sigma_\gamma = 10$.

We obtain posterior parameter samples by running three parallel chains of the HMC algorithm, implemented in Stan as discussed in Section 4.2.3. Each

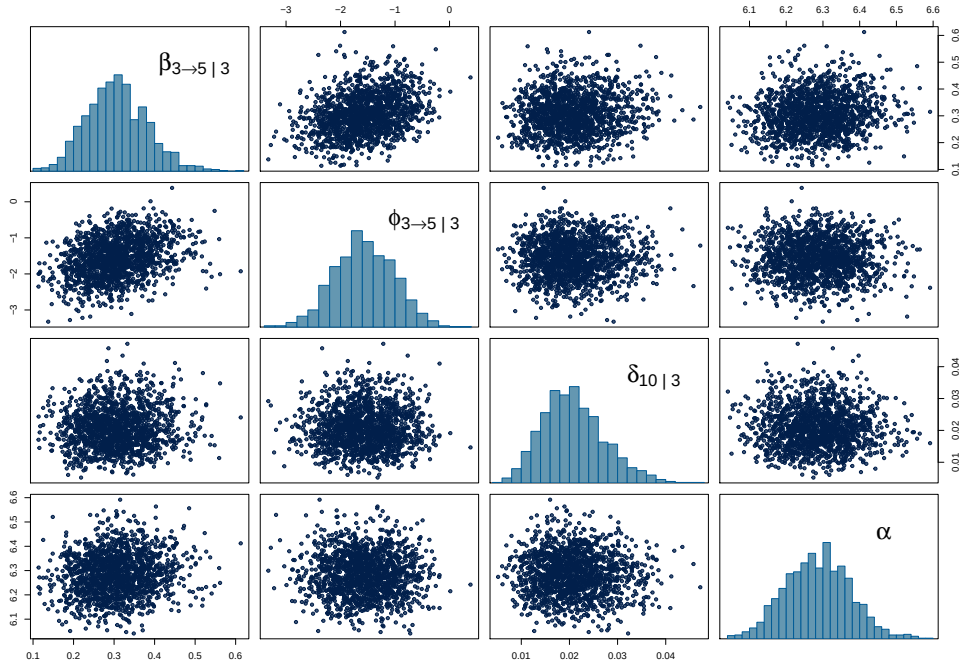


Figure 5.9: Pair-wise correlations with marginals along the diagonal for selected model parameters from the Matrix beta model for the marks in expression (5.8).

chain is initialised with different starting values and run for a total of 500 iterations post the warm-up phase. Figure 5.8 is a chain-wise trace plot of four model parameters and Table 5.20 gives the descriptive statistics along with the convergence diagnostics. All parameters had a potential scale reduction factor $\hat{R} < 1.1$ confirming the convergence of the sampling algorithm. The minimum effective sample size (N_{eff} column in Table 5.20) across all parameters is 330.4, greater than 20% of the total sample size, which indicates that the sampler returned samples with satisfactory autocorrelation. Figure 5.9 is a correlation plot for a selection of model parameters with the posterior marginals plotted along the diagonals and we do not observe any significant correlations between the parameters.

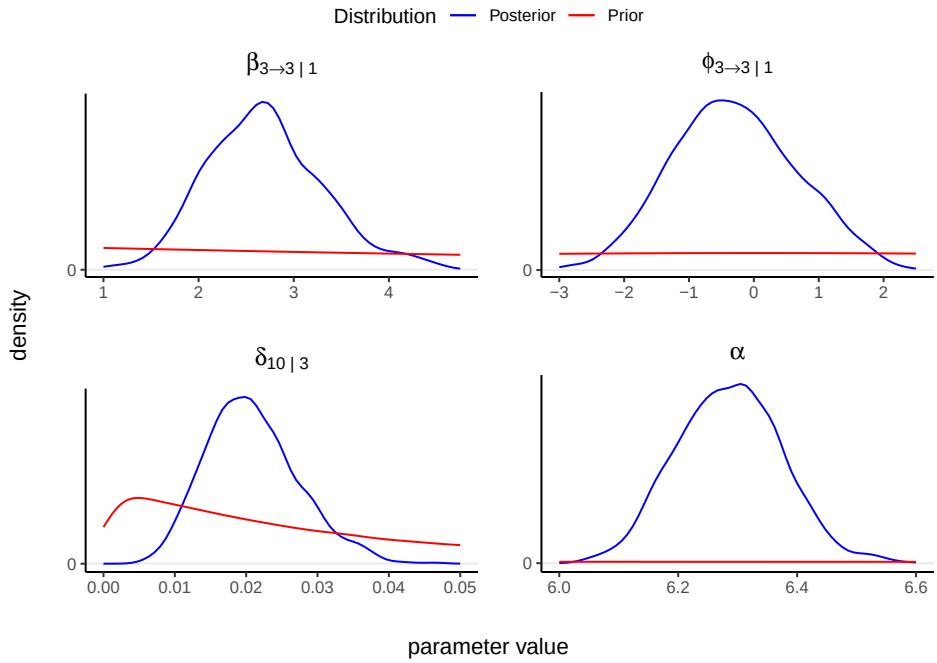


Figure 5.10: Visualising the impact of prior specifications by overlaying the posterior and prior densities for selected model parameters from the Matrix beta model for the marks in expression (5.8).

Impact of prior distributions

We assess the impact of the prior specifications for the Matrix beta model for the marks by comparing the variance of the parameter samples from their posterior and prior distributions as discussed in Section 4.1.4. Along with the individual variances, the ratio of the prior to posterior variance is provided in Table 5.21 for selected parameters.

The ratios are much larger than 1, indicating the posterior distributions of the parameters have concentrated around the maximum likelihood estimates. Figure 5.10 illustrates the flatness of the one-dimensional priors visually, by overlaying the marginal densities of the posterior and prior distributions.

parameter	posterior variance	prior variance	ratio of variance
$\beta_{3\to3\|1}$	0.4124	100	242.4543
$\varphi_{3\to3\|1}$	0.9677	100	103.3281
$\delta_{10\|3}$	4.16×10^{-5}	1.05×10^{-3}	25.2639
α	0.0078	100	12813.0524

Table 5.21: Quantifying the impact of prior specifications by computing the ratio of the prior to posterior variance for selected model parameters from the Matrix beta model for the marks in expression (5.8). Ratios much larger than 1 indicate that the prior distributions for the parameters are flat compared to their corresponding posterior distributions.

5.8 Model evaluation

In this section, we present results from the two approaches to evaluate the accuracy of the Bayesian models for point processes detailed in Section 4.4. Recall that all models were fitted using the first 40 game periods in the 2013/14 season as *training data* and for model evaluation, we use as *test data*, the 10 game periods immediately following the *training data*.

5.8.1 Log point-wise predictive density

The first approach for model evaluation, detailed in Section 4.4.1, relies on using the log-likelihood of the test data evaluated at the posterior parameter samples to compute a log score. For a single process in the *test data* with n observed events, the log point-wise predictive density can be computed using R posterior samples as

$$\widehat{lpd} = \sum_{i=1}^{n} \log \left(\frac{1}{R} \sum_{k=1}^{R} p\left(t_i, z_i, m_i \mid \boldsymbol{y}_k\right) \right) , \tag{5.14}$$

where $p\left(t_i, z_i, m_i \mid \boldsymbol{y}_k\right)$ is the likelihood of the i-th event in the process evaluated at the posterior sample $\boldsymbol{y}_k$.

Table 5.22 gives the model-wise log posterior densities $\widehat{lpd}$, cumulated over the 10 game periods in the *test data*. The number of estimated parameters

model	N_{par}	$\widehat{lpd}$
Homogeneous Poisson (Baseline)	90	-29062.95
Matrix beta ($W = 10$, $N = 50$)	538	-18205.61
Matrix beta ($W = 5$, $N = 50$)	538	-18067.13
Markov chain (Baseline)	870	-17815.66
Scalar beta	901	-17751.13
Vector beta	915	-17743.32
Matrix beta with teams ($W = 5$, $N = 100$)	1539	-17536.98
Matrix beta ($W = 10$, $N = 100$)	988	-17411.31
Matrix beta ($W = 5$, $N = 100$)	988	-17259.54

Table 5.22: Cumulative log posterior densities $\widehat{lpd}$ over 10 game periods in the test data for all fitted models along with the number of estimated parameters (N_{par}) in each model. For the Matrix beta models, W is the number of transient events and N is the number of significant event pairs identified in the rule-based framework for reducing model complexity.

(N_{par} column in Table 5.22) gives an indication of the complexity of each model. For the Matrix beta models, which used a rule-based framework for reducing model complexity, the tuning parameters W and N in the parenthesis are the number of transient events and the number of significant event pairs identified respectively.

The Matrix beta model with the setting ($W = 5$, $N = 100$) performs the best among the list of fitted models. Despite having comparable complexity with the Vector beta, the Scalar beta and the Markov chain based baseline models, the Matrix beta model ($W = 5$, $N = 100$) is able to outperform them significantly. The poorer performance of the the Matrix beta model with teams ($W = 5$, $N = 100$), compared to its counterpart without the team information, is due to over-fitting which is verified by fact that the model with team information does indeed perform the best in *training*. We believe the over-fitting can be reduced by training on a larger number of games. In the current experiment, each team plays just a single game each at their home and away venues. The size of the *training data* used in the modelling experiment was limited by the computational resources available

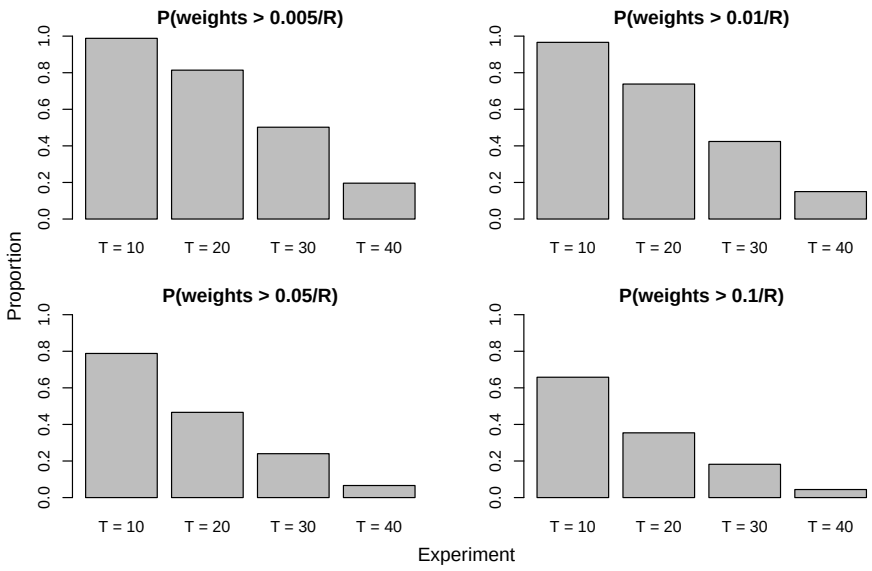


Figure 5.11: Proportion of importance sampling weights greater than a series of progressively increasing thresholds where $R = 500$ is number of posterior samples. The weights are calculated for the first game period in the test data given the event history up to time T in minutes.

at our disposal.

5.8.2 Simulation based validation

The second approach for model evaluation discussed in Section 4.4.2, works by simulating the sequence of events in a specified interval given its history, and then the simulated event counts are validated against the true observed counts. Figure 5.13 shows the set-up of the validation experiments, where we evaluate the models in four separate two-minute prediction intervals within each game period in the *test data* given the observed history of events before each interval.

Prior to simulation, the posterior samples are first updated given the observed history of events up to the prediction interval, using the importance sampling method in Section 4.3. For each posterior sample $\boldsymbol{y}_k$ we calculate its weight w_k as specified in expression (4.7), and then resample R times

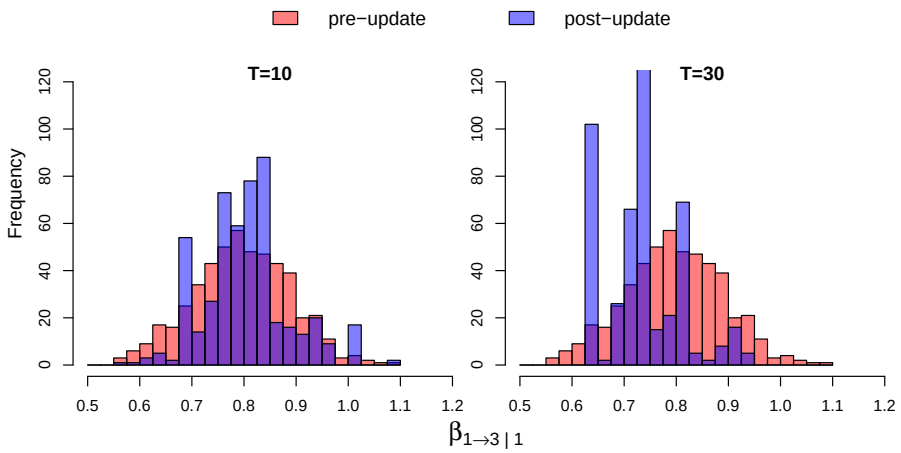


Figure 5.12: Posterior distributions for $\beta_{1\to 3|1}$ before and after updating given the event history up to time T of the first game period in the test data.

with replacement the samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_R$ with probabilities $w_1, \ldots, w_R$ to get the updated samples $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_R$. To inspect if the weights are imbalanced, we plot the proportion of weights greater than a series of progressively increasing thresholds in Figure 5.11.

The weights in Figure 5.11 are calculated for the first game period in the test data, based on four separate event histories corresponding to each experiment in Figure 5.13. For example, the weights corresponding to $T = 10$, are calculated based on the event history up to the 10th minute of the game period. We observe that the proportions in Figure 5.11 decrease from the first to the fourth experiment, indicating that the weights progressively become more concentrated. This is explained by the fact that there is increasingly more data to learn from.

Crucially, the proportions of the weights for a particular experiment appear to decrease gradually over the increasing thresholds (top-left to bottom right in Figure 5.11), showing no sign of significant imbalance. Figure 5.12 compares the posterior distributions for the parameter $\beta_{1\to 3|1}$ of the Matrix beta model, before and after updating based on two different event histories. As

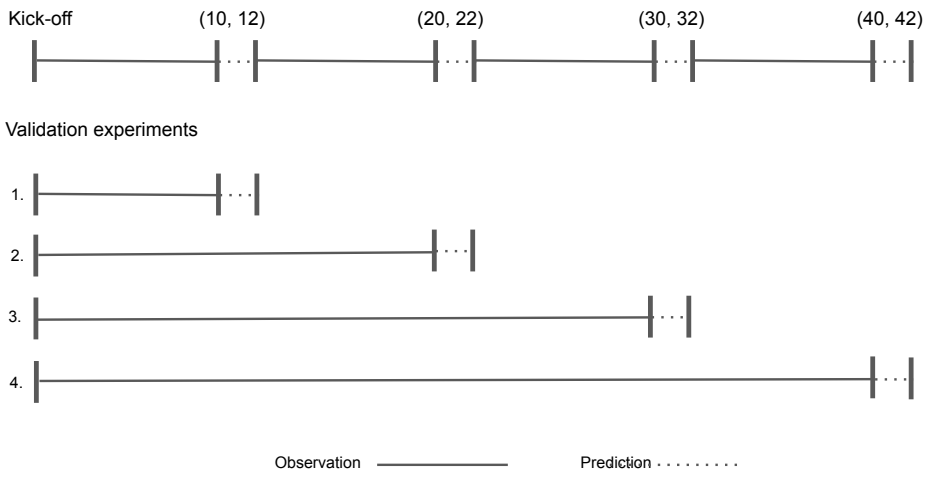


Figure 5.13: Design of the validation experiments, where the models are evaluated in four separate two-minute prediction intervals within each game period in the test data, given the observed history of events before each interval.

expected, we see a higher number of replications of the same samples in the $T = 30$ experiment as compared to the one at $T = 10$.

In the simulation based approach for validation, we evaluate the performance of a model by comparing the simulated counts of each event type against the true observed counts in a pre-defined prediction interval. We recognise that this not an optimal method to validate a point process model, as it does not take into account the inter-arrival times between the events or the order in which the events occur in the interval. Therefore, we recommend the model evaluation approach based on the log predictive density in Section 4.4.1 as the proper goodness of fit test.

However, mechanistic models, like those developed in this thesis, are often used to simulate real-world processes to predict the occurrence chance of a particular event in the near future. Keeping this in mind, we believe the series of validation experiments as designed in Figure 5.13, offers a reasonable framework to the test the predictions from our proposed model.

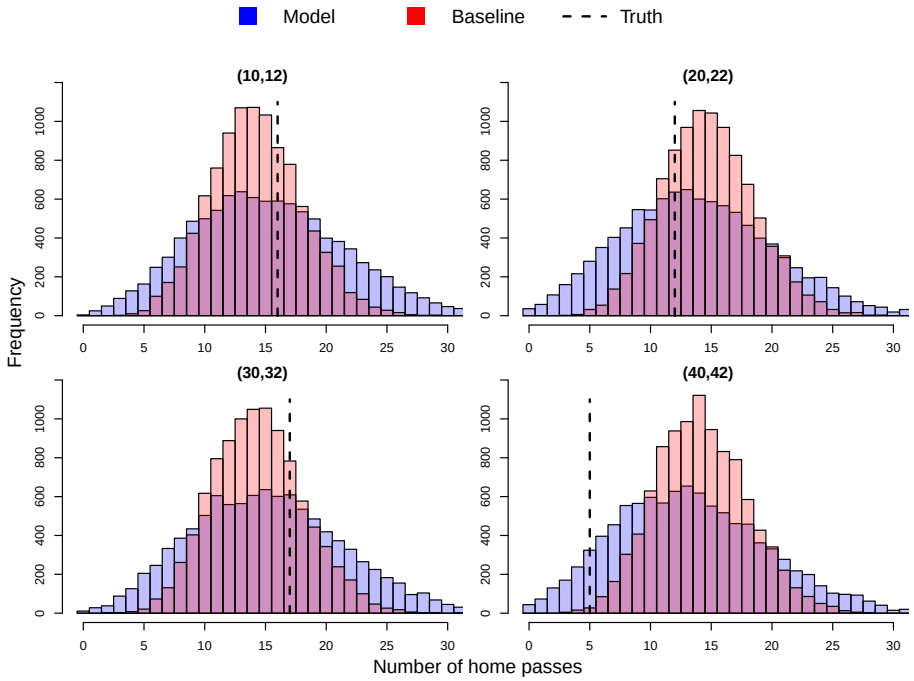


Figure 5.14: Predictive distributions of the number of successful passes by the home team (Arsenal) in four intervals of the first half in the game between Arsenal and Tottenham Hotspur in the test data. The observed count (truth) is given by the vertical dashed line.

As an example, in the validation experiments, we compare the best performing Matrix beta model with the setting ($W = 5$, $N = 100$) to the baseline homogeneous Poisson model. Henceforth in this section, we refer to the Matrix beta model simply as the 'model' and the homogeneous Poisson model as the 'baseline'. For each of the ten game periods in the test set, we follow the simulation framework proposed in Section 4.4.2 and simulate events in each prediction interval $Q = 500$ times for each of the $R = 500$ updated samples from the posterior.

Figure 5.14 shows the results from the validation experiments where we compare the performance of model and the baseline with respect to the observed count (truth) of a particular event type in the interval. In four separate 2-minute intervals of the first half in the game between Arsenal

Scoring Rule	(10, 12)	(20, 22)	(30, 32)	(40, 42)
Logarithmic	6	8	8	9
Brier	5	6	6	7
Spherical	5	6	6	7
Ranked Probability	5	8	6	7
Squared Error	5	8	7	7
Dawid Sebastiani	4	7	7	8

Table 5.23: Game periods (out of ten) where the model outperforms the baseline in each prediction interval of the validation experiments as designed in Figure 5.13.

and Tottenham Hotspur in the test set, for both the model and the baseline, we plot the distribution of the simulated successful passes by the home team (Arsenal). The baseline distribution is nearly identical across all intervals as expected, as the homogeneous Poisson process is memory-less. In the first (top-left in Figure 5.14) and third (bottom-left) intervals, we observe that the model distribution is shifted to the right and therefore predicting a higher count on average. In contrast, in the second (top-right) and fourth (bottom-right) intervals, the model distribution in shifted to the left and therefore predicting a lower count on average. In all scenarios, the model distributions agree well with the observed count and the model appears to outperform the baseline.

We formally evaluate the models using the performance measures defined in Section 4.4.3, that validate the predictive distribution for each event type in the interval against the observed truth. Table 5.23 presents the results of the validation experiments showing the number of instances out of the ten game periods in the test set, where the model outperforms the baseline. In Table 5.23, we aggregate the scores over all event types within each game period to arrive at a single score per model per experiment. We then count a success if the model achieves a lower score compared to the baseline. The model outperforms the baseline in 19 of the 24 score-interval combinations

Figure 5.15: Scoring rules for selected event types within a randomly chosen game in the test set with the prediction interval intervals along the x-axis.

and appears to get progressively better over the intervals. This is likely explained by the increasing amounts of data to learn from while updating the parameter samples.

Figure 5.15 shows the comparison of the scoring rules for selected event types from the validation experiments conducted for the first half in the game between Arsenal and Tottenham Hotspur in the test set. By definition, for all proper scoring rules, lower scores indicate better performance. It is interesting to see the consistency of the result across all scoring rules within each event type for any particular interval. For example, the Log score for the Home_Pass_S event (row 1 column 1 in Figure 5.15) in interval 1 (I1) shows the baseline having a lower score and outperforming the model. This agrees with all other scores for that event type in the same interval. For each scoring rule, the model outperforms the baseline in 10 of the 16 event-interval combinations, confirming the superior performance of the model.

5.9 Parameter description

Following on from the inference for the Matrix beta model with team information presented in Section 5.7.5, in this section, we look at the estimated parameters of the excitation based model for the marks in detail. We build on the intuition for the model parameters we had developed in Section 3.3 to interpret the parameters in the context of football and illustrate the potential of the model to provide insights into the underlying dynamics of the game.

5.9.1 Background mark probability

The background mark probability $\delta_{m|z}$ in expression (5.8) is the probability an event that occurs in location z has a mark m, if the event is triggered solely by the background component. Table 5.24 gives the posterior means of the $\delta_{m|z}$'s and we observe that the background mark probabilities for the home and away teams nearly mirror each other across the zones. This is explained by fact that the attacking zone for the home team is the defensive zone for the away team and vice-versa. The lack of variation between the home and away parameters implies that the background component of the game is not influenced by the *home advantage* effect. We also observe that the successful Pass events account for the majority of the background probability mass, while events like Shots and Goals have nearly 0 mass. This suggests that the Shot and Goal events are unlikely to originate solely from the background component, but rather are triggered by the excitation from a previous event.

5.9.2 Excitation factor

The excitation factor α in expression (5.8) is a scaling factor applied to the contributions from the previous occurrences to the event mark probability. The posterior mean of α is 6.28, which indicates that event sequences in foot-

mark label	1	2	3	mark label	1	2	3
Home_Win	.	0.01	.	Away_Win	.	0.01	.
Home_Dribble	0.03	0.02	0.01	Away_Dribble	.	0.02	0.03
Home_Pass_S	0.56	0.23	0.03	Away_Pass_S	0.03	0.11	0.59
Home_Pass_U	0.06	0.04	0.03	Away_Pass_U	0.02	0.07	0.05
Home_Shot	.	.	.	Away_Shot	.	0.01	.
Home_Keeper	0.04	.	.	Away_Keeper	.	.	0.05
Home_Save	.	.	.	Away_Save	.	.	.
Home_Clear	0.05	0.01	.	Away_Clear	0.01	0.02	0.01
Home_Lose	0.03	0.04	0.01	Away_Lose	0.01	0.07	0.06
Home_Goal	.	.	0.02	Away_Goal	0.02	.	.
Home_Foul	0.03	0.08	0.02	Away_Foul	0.03	0.06	0.02
Home_Out_Throw	.	0.02	.	Away_Out_Throw	.	.	.
Home_Out_GK	.	.	.	Away_Out_GK	.	.	.
Home_Out_Corner	.	.	.	Away_Out_Corner	.	.	.
Home_Pass_O	0.01	0.08	0.03	Away_Pass_O	0.02	0.05	0.01

Table 5.24: Posterior means of the zone dependent background mark probabilities $\delta_{m|z}$ for $z \in \{1,2,3\}$ from the Matrix beta model for the marks in expression (5.8). The dots $(\cdot)$ denote $\delta_{m|z}$ values less than 0.01.

ball have a strong dependence on their history, and the contributions from previous occurrences are weighted approximately $\exp(6.28) \approx 533$ times greater in comparison to the background component.

5.9.3 Decay rates

The decay rate $\beta_{m_j \to m_i|z_i}$ in expression (5.8) is the exponential decay rate of the excitation caused by an event of mark m_j on an event of mark m_i in the location z_i. By allowing the decay rates to depend on the pair of marks involved in the excitation, we had hoped to account for scenarios like a Corner event exciting a Pass_S event in the short term and a Shot event in the longer term. Indeed, the estimated posterior means $\beta_{\text{Corner}\to\text{Pass\_S}} = 1.76$ and $\beta_{\text{Corner}\to\text{Shot}} = 0.31$ confirm that the Corner $\to$ Shot excitation decays at a much slower rate compared to the Corner $\to$ Pass_S excitation. Note that for notational convenience, we ignore the location and team information in this section and present the posterior means aggregated across the home and

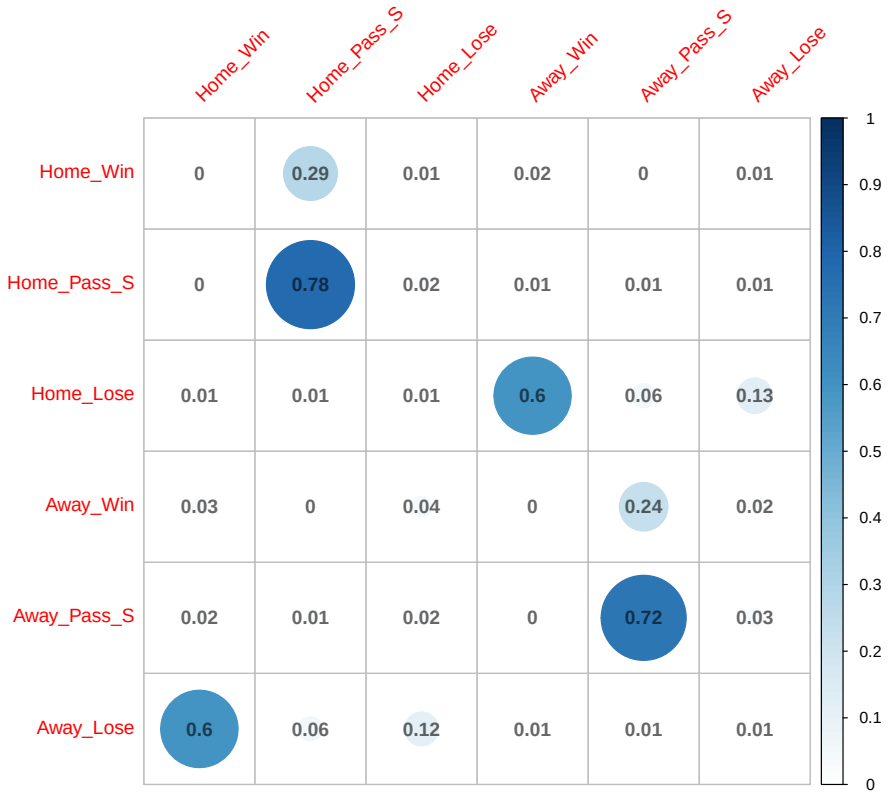


Figure 5.16: Posterior means of the event conversion probabilities $\gamma_{m_j \to m_i|z_i}$ for a selection of event pairs corresponding to the location $z = 2$ from the Matrix beta model for the marks in expression (5.8). The $\gamma_{m_j \to m_i|z_i}$'s are computed for a hypothetical match-up where the baseline team West Ham United is chosen as both the home as well as the away team to negate the impact of team abilities.

away teams at their corresponding attacking locations.

5.9.4 Conversion rates

The parameter $\gamma_{m_j \to m_i|z_i}$ in expression (5.8) is the probability the excitation from an event of mark m_j triggers an event of mark m_i in the location z_i. Figure 5.16 gives the posterior means of $\gamma_{m_j \to m_i|z_i}$'s for a selection of event pairs corresponding to the midfield region ($z = 2$). Team information is incorporated into the event conversion rates using the baseline logit specifi-

cation in expression (5.9), where the baseline team is chosen to be West Ham United (assigned an ability 0). The $\gamma_{m_j \to m_i|z_i}$'s in Figure 5.16 are computed for a hypothetical match-up where the baseline team West Ham United is chosen as both the home as well as the away team to negate the impact of team abilities.

The probabilities for the Home_Win $\to$ Home_Pass_S and Home_Pass_S $\to$ Home_Pass_S conversions are higher compared to their away team counterparts, indicating that the home team is superior in retaining possession of the ball. In this way, we not only confirm the well-known *home advantage* effect, but also quantify it with the claim that the home team is approximately 6% more likely to retain possession of the ball in the midfield region.

5.9.5 Team ability

The mark dependent team ability parameters $\omega_{h,m}$ in expression (5.9) capture the relative ability of a team to make an event conversion in comparison to a baseline team. The baseline team is assigned an ability value of 0 and an $\omega_{h,m} > 0$ indicates that for the team h, a previous event is more likely to trigger an event of mark m when compared to the baseline team. Being the last team when listed alphabetically, West Ham United is taken as the baseline in the modelling experiment.

The posterior means of the team ability parameters in Table 5.25 show the relative abilities $\omega_{h,\text{Home\_Pass\_S}}$ and $\omega_{h,\text{Away\_Pass\_S}}$ of a team h to complete a successful pass and retain possession when playing at their home and away venues respectively. We obtain a ranking for the teams based on their relative passing ability by ordering the posterior means of the cumulative ability ($\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$) in Table 5.25.

Figure 5.17 provides a ridge-line plot of the posterior distribution of the parameters $\omega_{h,\text{Home\_Pass\_S}}$ and $\omega_{h,\text{Away\_Pass\_S}}$. Once again, the teams are listed in

Team	$\omega_{h,\text{Home\_Pass\_S}}$	$\omega_{h,\text{Away\_Pass\_S}}$
Manchester City	0.70	0.96
Chelsea	0.71	0.78
Arsenal	0.57	0.80
Southampton	0.75	0.54
Manchester United	0.79	0.46
Everton	0.51	0.73
Liverpool	0.65	0.56
Hull City	0.57	0.51
Tottenham Hotspur	0.25	0.81
Fulham	0.68	0.35
Stoke City	0.41	0.54
Newcastle United	0.56	0.34
Sunderland	0.70	0.03
Swansea City	0.46	0.24
Cardiff City	0.23	0.40
Norwich City	-0.09	0.72
Crystal Palace	0.27	0.28
West Bromwich Albion	0.21	0.22
Aston Villa	0.37	-0.11
West Ham United	0.00	0.00

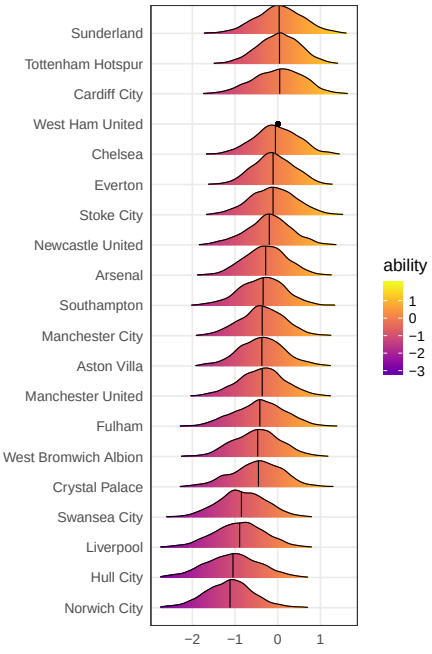
Table 5.25: Posterior means of team ability parameters ordered by the cumulative ability ($\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$) of the team h to complete a successful pass and retain possession. Team information is incorporated into the event conversion rates using the baseline logit specification in expression (5.9).

the decreasing order of the means of their respective posterior distributions which are marked in the figure by vertical lines. We observe that Manchester United, the team with the highest ability to retain possession in home games (Figure 5.17a), drop significantly down in the rankings for the away games (Figure 5.17b). This suggests that Manchester United might be adopting a more direct, counter-attacking playing style in their away games compared to a possession based approach in the home games.

Figure 5.18a provides a ridge-line plot of the posterior distribution of the cumulative ability of a team to attempt a shot on goal. A higher $\omega_{h,\text{Home\_Shot}}$, for example, indicates that for the team h, an event like Home_Pass_S is more likely to trigger a Home_Shot. We do not expect the cumulative abilities

Figure 5.17: Posterior distribution of the parameters $\omega_{h,\text{Home\_Pass\_S}}$ in (a) and $\omega_{h,\text{Away\_Pass\_S}}$ in (b), from the baseline logit specification for incorporating team abilities in expression (5.9). Teams are ranked in the decreasing order of the means of their respective posterior distributions shown by the overlayed vertical lines.

$\omega_{h,\text{Home\_Shot}} + \omega_{h,\text{Away\_Shot}}$ of the dominant teams to be high, as they might prefer to make additional passes to create better goal scoring opportunities. A weaker team, on the other hand, typically has fewer opportunities to attack and therefore, is more likely to attempt a shot on goal when possible. Indeed, this is what we observe in Figure 5.18, where we compare the team rankings based on their cumulative ability $\omega_{h,\text{Home\_Shot}} + \omega_{h,\text{Away\_Shot}}$ with the number of shots per pass completed in the attacking third (S/P column in Figure 5.18b) in the training data. The comparison between Cardiff City and Norwich City is an interesting example of two teams that appear to be similar with 18 and 19 shots on goal attempted, respectively, in their two games in the training data. However, the two teams are at the opposite ends of the ranking based on their cumulative ability $\omega_{h,\text{Home\_Shot}} + \omega_{h,\text{Away\_Shot}}$,



(a)

Team	Shots	Passes	S/P
Cardiff City	18	115	0.16
Sunderland	28	195	0.14
Tottenham Hotspur	37	261	0.14
Aston Villa	20	153	0.13
Newcastle United	22	180	0.12
Crystal Palace	18	148	0.12
Stoke City	24	199	0.12
Everton	40	342	0.12
Chelsea	29	248	0.12
Fulham	19	163	0.12
West Ham United	22	194	0.11
West Bromwich Albion	18	159	0.11
Swansea City	23	214	0.11
Arsenal	30	280	0.11
Southampton	23	223	0.10
Liverpool	28	292	0.10
Manchester United	22	238	0.09
Manchester City	30	330	0.09
Norwich City	19	227	0.08
Hull City	13	191	0.07

(b)

Figure 5.18: (a) Posterior distribution of $\omega_{h,\text{Home\_Shot}} + \omega_{h,\text{Away\_Shot}}$, the cumulative ability of a team h, relative to West Ham (baseline), to attempt a shot on goal. (b) The number of shots, passes completed in the attacking third and shots per pass completed in the attacking third (S/P) for each team in the training data.

capturing the clear difference between their attacking styles.

Figure 5.19a shows the team rankings based on the cumulative ability to trigger 5 different event types. For example, the Pass column ranks teams in the decreasing order of their posterior means of $\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$. The teams are ordered in Figure 5.19a by the rankings based on their cumulative passing ability. Despite training on just the first 20 out of 380 games of the 2013/14 season, the rankings based on the passing ability is a good predictor of the positions the teams finished in the final league table of the 2013/14 season in Figure 5.19b. We believe that by training on a larger number of games and ranking teams by event type, would not only result in a better predictor of team performance, but also provide valuable insight into

Team	Pass	Shot	Goal	Win	Save
Manchester City	1	11	1	15	11
Chelsea	2	5	11	20	4
Arsenal	3	9	3	5	7
Southampton	4	10	8	7	19
Manchester United	5	13	4	2	18
Everton	6	6	17	11	14
Liverpool	7	18	12	9	5
Hull City	8	19	15	1	10
Tottenham Hotspur	9	2	14	3	6
Fulham	10	14	7	14	3
Stoke City	11	7	9	12	8
Newcastle United	12	8	19	16	17
Sunderland	13	1	13	8	2
Swansea City	14	17	18	17	12
Cardiff City	15	3	2	19	15
Norwich City	16	20	10	4	20
Crystal Palace	17	16	16	13	16
West Bromwich Albion	18	15	20	10	1
Aston Villa	19	12	5	6	13
West Ham United	20	4	6	18	9

(a)

(b)

Figure 5.19: (a) Team rankings based on the cumulative ability to trigger a particular event type. For example, the column Pass, ranks teams in the decreasing order of their respective posterior means of $\omega_{h,\text{Home\_Pass\_S}} + \omega_{h,\text{Away\_Pass\_S}}$. (b) The final positions of the teams in the league table of the 2013/14 season taken from www.whoscored.com. The team rankings estimated using the cumulative passing ability in (a) is the best predictor of the final positions in the league table in (b).

the playing styles of the different teams.

5.10 Recovering hidden structure

The probability mass function for the marks in expression (5.8) does not directly provide any intuition towards the causality of the event occurrences. The mark probability of any event is determined by the combined additive effect from the background component and all previous occurrences. The only exception is for the first event in the sequence that is triggered solely from the background component. However, we may want to assume a causal

constraint that any event is triggered by exactly one of the previous events or the background and as this triggering is unobserved, we wish to recover this hidden branching structure.

The branching structure denoted by u_i, defined in Section 2.2.3, indicates whether the i-th event is an immigrant ($u_i = 0$) or an offspring of a previous event with index j ($u_i = j$). Given an observed event sequence $\mathcal{F}_T$, the conditional branching structure probabilities $\mathbb{P}(u_i = j \mid \mathcal{F}_{t_i})$ based on the model specification in expression (5.8) are

$$\mathbb{P}(u_i = 0 \mid \mathcal{F}_{t_i}) = \frac{\delta_{m_i|z_i}}{\delta_{m_i|z_i} + \sum_{t_k < t_i} \mathrm{e}^{\alpha - \beta_{m_k \to m_i|z_i}(t_i - t_k)} \gamma_{m_k \to m_i|z_i}},$$

$$\mathbb{P}(u_i = j \mid \mathcal{F}_{t_i}) = \begin{cases} \dfrac{\mathrm{e}^{\alpha - \beta_{m_j \to m_i|z_i}(t_i - t_j)} \gamma_{m_j \to m_i|z_i}}{\delta_{m_i} + \sum_{t_k < t_i} \mathrm{e}^{\alpha - \beta_{m_k \to m_i|z_i}(t_i - t_k)} \gamma_{m_k \to m_i|z_i}} & \text{for } t_j < t_i \\ 0 & \text{for } t_j \geq t_i \end{cases}. \quad (5.15)$$

Even if the underlying process does not allow a constraint where the events are triggered by exactly one of the previous events or the background, the branching structure probabilities in expression (5.15) quantify the relative contributions of the background and previous occurrences in the mark probability of the i-th event. Figure 5.20 shows the branching structure probabilities for all events in the first four minutes of the game between Chelsea and Hull City in the 2013/14 season. To illustrate the flexibility of the model to account for dependence between events over arbitrary durations of time, we highlight the event Home_Shot which has a higher probability of being an offspring of the event Home_Out_Corner than being an offspring of the more recent Home_Pass_S event.

5.11 Real-time simulation

Finally, we illustrate how the mechanistic modelling framework developed in this thesis can be used to simulate event sequences in football and ob-

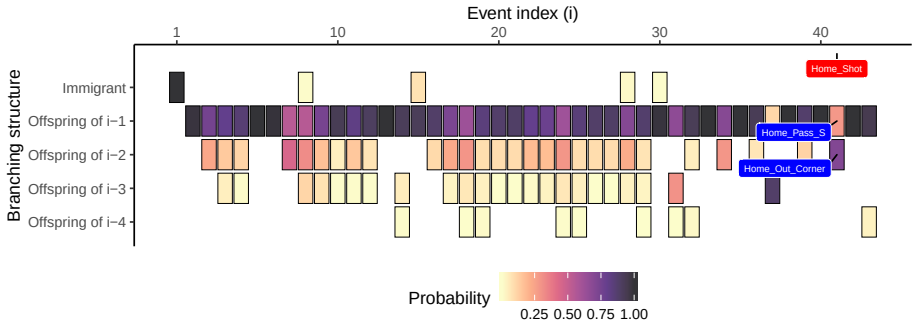


Figure 5.20: Branching structure probabilities for events in the first 4 minutes of the game between Chelsea and Hull City in the 2013/14 season of the English Premier League. The highlighted event Home_Shot has a higher probability of being an offspring of the event Home_Out_Corner than being an offspring of the more recent Home_Pass_S event.

tain predictions of event probabilities in real-time. We split the first half of the game between Arsenal and Tottenham Hotspur in the test data into 1-minute intervals starting at time 0 before any events have occurred. For each interval, given the history of events up to but not including the interval, we simulate events over the next one minute $Q = 100$ times for each of the $R = 500$ posterior samples after updating from the excitation-based Matrix beta model in Section 5.7.5 with the tuning parameter setting ($W = 5$, $N = 100$).

In Figure 5.21, we plot the proportion of all simulations within each interval where at least one Home Shot event was simulated, and use dotted lines to denote the intervals where an Home Shot event was actually observed. A quick inspection reveals that in 6 of the 8 intervals in which a Home_Shot is observed, the model predicts a probability greater than 0.26 (mean predicted probability over the 45 intervals). We believe the predictive performance can be improved by training on a larger number of games to get better estimates of the parameters capturing the team abilities as well as the underlying game dynamics. Also, the association rule framework for identifying significant event pairs in Section 5.6 can be tailored to event sequences in football.

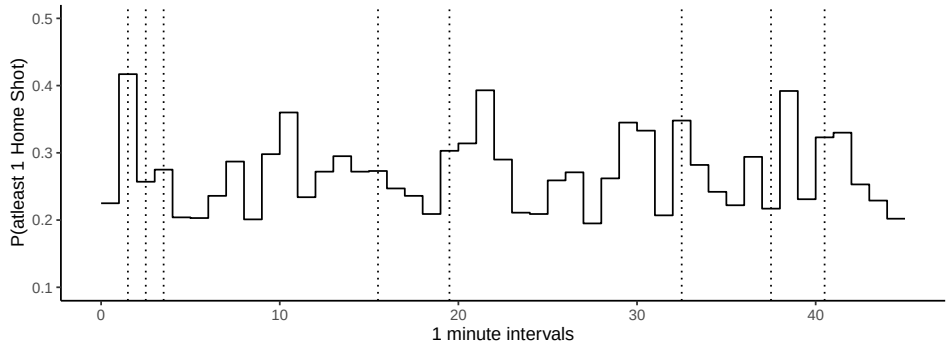


Figure 5.21: Forecasting the probability of observing at least one Home Shot event in 1-minute intervals over the first half of the game between Arsenal and Tottenham Hotspur in the test data. Intervals with observed Home Shot events are highlighted using dotted lines.

Events like shots and goals could be treated differently, as they are arguably the most interesting and certainly the most consequential in the outcome of football matches.

Chapter 6

Concluding remarks

Analysing event sequences using mechanistic point process models has the potential to describe the underlying scientific process that generated the data. However, such mechanistic models are typically specified using a joint conditional intensity and in general are not flexible enough to be applied to real-world datasets. It is also common to make strong restrictive assumptions like separability to simplify the model as the joint modelling of the components of the process can be challenging. In this thesis, we developed a flexible mechanistic modelling framework for marked spatio-temporal point processes that are suitable for a wide-range of applications.

The focus area of our work was motivated by the problem of modelling event sequences in association football. Most analyses in football are typically done manually by watching video footage or using simple frequency analysis of match events. Sophisticated analyses of football data, however, is mathematically challenging due to the continuous interaction between players within and across the two teams. We applied the flexible framework for marked spatio-temporal point processes to the event sequences in football with the aim of describing the game dynamics.

Building on the decomposition of a multivariate density function, in Chapter 3, we showed how the joint modelling in classical point process models, like

Hawkes processes, can be decoupled. We developed a flexible modelling framework that can, for example, retain the characteristic property of excitation in Hawkes processes, in the model for the marks while avoiding the clustering of event times. We provided details on the parameter estimation for such flexible processes via an EM (Expectation-Maximisation) algorithm that takes advantage of the inherent branching structure. In Chapter 4, we developed a comprehensive Bayesian approach for the modelling of flexible marked spatio-temporal point processes that can be readily applied to other applications, especially when on-line inference is necessary. We discussed a formal approach to evaluate the goodness of fit of Bayesian models using the out-of-sample log predictive density. We also developed a framework for updating the model parameters based on new data and then simulating a sequence of events, that can be implemented efficiently to make predictions in real-time.

In Chapter 5, we presented a case study on the application of mechanistic point process models to event sequences in football. We discussed in detail how the flexible modelling framework developed in this thesis can be tailored to separately model the components of the events in football, namely, the times, the locations and the event types. We were also able to incorporate team information into the model in a direct way that captures the relative abilities of the teams. We developed a method based on association rules to reduce the increased model complexity introduced by model extensions. The rule-based approach identifies significant event interactions within sequences in a clever way by placing thresholds on some measures of significance. We then evaluated the accuracy of the excitation based models by comparing against two baseline models and confirmed the superior performance of the models with excitation effects.

We provided a detailed parameter description showing how the model parameters can be used to gain valuable insight into football. The excitation

framework of the best performing model captures both the magnitudes and the durations of all pairwise event interactions across different locations. From the conversion rate parameters, we were able to quantify the well-known *home advantage* effect, and learned that the home team is approximately 6% more likely to retain possession of the ball in the midfield region. We also discussed how the team ability parameters can be used to obtain rankings for the teams by event type, that can be used as predictors for team performance. The team ability parameters also captured some interesting differences in the playing styles of the teams, that weren't immediately apparent. In this way, the model along with its parameters can be used to develop a deeper understanding of the game-play by the coaching staff and inform strategic decision making. The proposed model can also be used to simulate the sequence of events in a game to obtain real-time predictions of event probabilities. We believe these predictions would enhance, among others, the viewing experience of televised games.

We also identified some key ideas for future work and extensions. Our modelling framework is built on the traditional Hawkes process model that could be extended to incorporate non-linear excitation effects as well as inhibitory effects. Also, as we do not account for the evolution of the parameters over games, having a time-varying hierarchical structure could prove to be a valuable addition. The use of a more constrained prior specification to handle model complexity would also be an interesting area to explore. Finally, we also feel it would be worthwhile to adopt a more involved on-line inference algorithm like the SMC method to maximise performance.

To conclude, we believe the flexible modelling framework developed in this thesis could be a valuable addition to the arsenal of point process modellers, aiding in the development of customised models borrowing the required characteristics from existing models. And the modelling of event sequences in football using excitation based mechanistic models offer the capability to

describe event interactions, evaluate team performance and forecast events that can make a significant impact in a highly competitive sport.

Bibliography

Agrawal, R., T. Imieliundefinedski, and A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2), 207–216.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Ahrens, J. H. and U. Dieter (1982). Generating gamma variates by a modified rejection technique. *Communications of the ACM 25*(1), 47–54.

Bernardo, J. and A. Smith (2007). *Bayesian Theory*. New York: John Wiley & Sons.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. arXiv:1701.02434.

Borrie, A., G. K. Jonsson, and M. S. Magnusson (2002). Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. *Journal of sports sciences 20*(10), 845–852.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics 141*(2), 876–912.

Brémaud, P. and L. Massoulié (1996). Stability of nonlinear hawkes processes. *The Annals of Probability 24*(3), 1563–1588.

Brin, S., R. Motwani, J. D. Ullman, and S. Tsur (1997). Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec. 26*(2), 255–264.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika 89*(3), 539–552.

Clemente, F. M., M. S. Couceiro, F. M. L. Martins, and R. S. Mendes (2015). Using network metrics in soccer: a macro-analysis. *Journal of human kinetics 45*(1), 123–134.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I* (2nd ed.). New York: Springer-Verlag.

Decroos, T., L. Bransen, J. Van Haaren, and J. Davis (2018). Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. arXiv:1802.07127.

Decroos, T., V. Dzyuba, J. V. Haaren, and J. Davis (2017). Predicting soccer highlights from spatio-temporal match event streams. In S. P. Singh and S. Markovitch (Eds.), *AAAI*, pp. 1302–1308.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B: Methodological 39*(1), 1–22.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (3rd ed.). Boca Raton, Florida: CRC Press.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Duch, J., J. S. Waitzman, and L. A. N. Amaral (2010). Quantifying the performance of individual players in a team activity. *PLOS One 5*(6), e10937+.

Gabry, J. and T. Mahr (2019). bayesplot: Plotting for bayesian models. R package version 1.7.1.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, Florida: CRC Press.

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence 6*, 721–741.

Geyer, C. J. (2011). Introduction to markov chain monte carlo. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 3–48. Chapman & Hall / CRC Press.

González, J. A., F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu (2016). Spatio-temporal point process statistics: a review. *Spatial Statistics 18*, 505–544.

Grund, T. U. (2012). Network structure and team performance: The case of english premier league soccer teams. *Social Networks 34*(4), 682–690.

Gudmundsson, J. and M. Horton (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR) 50*(2), 1–21.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*(1), 83–90.

Hawkes, A. G. and D. Oakes (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability 11*(03), 493–503.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Ishwaran, H., J. S. Rao, et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics 33*(2), 730–773.

Kahn, H. and T. E. Harris (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series 12*, 27–30.

Kingman, J. F. C. (1993). *Poisson processes*. New York: Clarendon Press.

Kleinrock, L. (1975). *Queueing systems. Volume I: theory*. New York: John Wiley & Sons.

Kong, A. (1992). A note on importance sampling using standardized weights. Technical report, Dept. of Statistics, University of Chicago.

Laub, P. J., T. Taimre, and P. K. Pollett (2015). Hawkes processes. arXiv:1507.02822.

Lewis, P. (1969). Asymptotic properties and equilibrium conditions for branching poisson processes. *Journal of Applied Probability 6*(2), 355—-371.

Lewis, P. W. and G. S. Shedler (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly 26*(3), 403–413.

Mackay, N. (2017). Predicting goal probabilities for possessions in football. Master's thesis, Vrije Universiteit Amsterdam.

Martin, A. D., K. M. Quinn, and J. H. Park (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software 42*(9), 1–21.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association 46*(253), 68–78.

Mei, H. and J. M. Eisner (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pp. 6754–6764.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Millar, R. B. (2004). Sensitivity of bayes estimators to hyper-parameters with an application to maximum yield from fisheries. *Biometrics 60*(2), 536–542.

Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association 106*(493), 100–108.

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto Toronto, ON, Canada.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 62–116. Chapman & Hall / CRC Press.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical programming 120*(1), 221–259.

Norris, J. R. (1997). *Markov Chains*. Cambridge: Cambridge University Press.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory 27*(1), 23–31.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics 50*(2), 379–402.

Passos, P., K. Davids, D. Araújo, N. Paz, J. Minguéns, and J. Mendes (2011). Networks as a novel tool for studying team ball sports as complex social systems. *Journal of Science and Medicine in Sport 14*(2), 170–176.

Pena, J. L. and H. Touchette (2012). A network theory analysis of football strategies. arXiv:1206.6904.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raiffa, H. and R. Schlaifer (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability 15*(3), 623–642.

Robberechts, P., J. Van Haaren, and J. Davis (2019). Who will win it? an in-game win probability model for football. arXiv:1906.05029.

Routley, K. and O. Schulte (2015). A markov game model for valuing player actions in ice hockey. In M. Meila and T. Heskes (Eds.), *UAI*, pp. 782–791.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*. London: Chapman & Hall.

Stan Development Team (2020). Cmdstan: the command-line interface to stan. Version 2.22.1.

Van Haaren, J., S. Hannosset, and J. Davis (2016). Strategy discovery in professional soccer match data. In *Proceedings of the KDD-16 Workshop on Large-Scale Sports Analytics*, pp. 1–4.

Veen, A. and F. P. Schoenberg (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association 103*(482), 614–624.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing 27*(5), 1413–1432.

Wang, Q., H. Zhu, W. Hu, Z. Shen, and Y. Yao (2015). Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2197–2206.